DOCUMENT RESUME

ED 038 306	24	SE 008 253

AUTHOR	Kriewall, Thomas E.
TITLE	Applications of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction, Part 1.
INSTITUTION	Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
REPORT NO	TR-103
BUREAU NO	BR-5-0216
PUB DATE	Oct 69
CONTRACT	OEC-5-10-154
NOTE	191p.
EDRS PRICE	EDRS Price MF-$0.75 HC-$9.65
DESCRIPTORS	*Computer Assisted Instruction, Doctoral Theses, *Individualized Instruction, *Information Theory, *Learning, Mathematics Education, *Research

ABSTRACT

This Technical Report is concerned with various problems of instructional management encountered in situations which stress self-selection and self-pacing principles. These problems deal primarily with the efficient utilization and allocation of human and material resource materials to formulate an operational, individualized, inquiry-learning environment. The specific purpose of this report was to develop a decision system based on a relevant theory of criterion-referenced tests and explore its potential for solving some problems which arise in the management of individualized mathematics curricula. The Report is a part of the Project on Computer-Managed Systems of Mathematics Instruction. (FL)

BR 5-0216
PA-24
SE

ED038306

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

No. 103

APPLICATIONS OF INFORMATION THEORY AND ACCEPTANCE SAMPLING PRINCIPLES TO THE MANAGEMENT OF MATHEMATICS INSTRUCTION

PART I

Report from the Project on
Computer-Managed Systems of Mathematics Instruction

SE 008 253

U.S. Office of Education Center No. C-03 Contract OE 5-10-154

MAR 3 0 1970

ED038306

Technical Report No. 103

APPLICATIONS OF INFORMATION THEORY AND ACCEPTANCE SAMPLING PRINCIPLES TO THE MANAGEMENT OF MATHEMATICS INSTRUCTION

PART I

Report from the Project on
Computer-Managed Systems of Mathematics Instruction

By Thomas E. Kriewall, Project Director

M. Vere DeVault, Professor of Curriculum and Instruction,
Chairman of the Examining Committee, and Principal Investigator

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

October 1969

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

Center No. C-03 / Contract OE 5-10-154

NATIONAL EVALUATION COMMITTEE

Samuel Brownell
Professor of Urban Education
Graduate School
Yale University

Launor F. Carter
Senior Vice President on
Technology and Development
System Development Corporation

Francis S. Chase
Professor
Department of Education
University of Chicago

Henry Chauncey
President
Educational Testing Service

Martin Deutsch
Director, Institute for
Developmental Studies
New York Medical College

Jack Edling
Director, Teaching Research
Division
Oregon State System of Higher
Education

Elizabeth Koontz
President
National Education Association

Roderick McPhee
President
Punahou School, Honolulu

G. Wesley Sowards
Director, Elementary Education
Florida State University

Patrick Suppes
Professor
Department of Mathematics
Stanford University

***Benton J. Underwood**
Professor
Department of Psychology
Northwestern University

UNIVERSITY POLICY REVIEW BOARD

Leonard Berkowitz
Chairman
Department of Psychology

Archie A. Buchmiller
Deputy State Superintendent
Department of Public Instruction

***James W. Cleary**
Vice Chancellor for Academic
Affairs

Leon D. Epstein
Dean
College of Letters and Science

John Guy Fowlkes
Director
Wisconsin Improvement Program

Robert E. Grinder
Chairman
Department of Educational
Psychology

H. Clifton Hutchins
Chairman
Department of Curriculum and
Instruction

Clauston Jenkins
Assistant Director
Coordinating Committee for
Higher Education

Herbert J. Klausmeier
Director, R & D Center
Professor of Educational
Psychology

Donald J. McCarty
Dean
School of Education

Ira Sharkansky
Associate Professor of Political
Science

Henry C. Weinlick
Executive Secretary
Wisconsin Education Association

M. Crawford Young
Associate Dean
The Graduate School

EXECUTIVE COMMITTEE

Edgar F. Borgatta
Brittingham Professor of
Sociology

Max R. Goodson
Professor of Educational Policy
Studies

Russell J. Hosler
Professor of Curriculum and
Instruction and of Business

***Herbert J. Klausmeier**
Director, R & D Center
Professor of Educational
Psychology

Wayne Otto
Professor of Curriculum and
Instruction (Reading)

Robert G. Petzold
Associate Dean of the School
of Education
Professor of Curriculum and
Instruction and of Music

Richard L. Venezky
Assistant Professor of English
and of Computer Sciences

FACULTY OF PRINCIPAL INVESTIGATORS

Ronald R. Allen
Associate Professor of Speech
and of Curriculum and
Instruction

Vernon L. Allen
Associate Professor of Psychology
(On leave 1968-69)

Nathan S. Blount
Associate Professor of English
and of Curriculum and
Instruction

Robert C. Calfee
Associate Professor of Psychology

Robert E. Davidson
Assistant Professor of
Educational Psychology

Gary A. Davis
Associate Professor of
Educational Psychology

M. Vere DeVault
Professor of Curriculum and
Instruction (Mathematics)

Frank H. Farley
Assistant Professor of
Educational Psychology

John Guy Fowlkes (Advisor)
Professor of Educational
Administration
Director of the Wisconsin
Improvement Program

Lester S. Golub
Lecturer in Curriculum and
Instruction and in English

Max R. Goodson
Professor of Educational Policy
Studies

Warren O. Hagstrom
Professor of Sociology

John G. Harvey
Associate Professor of
Mathematics and Curriculum
and Instruction

Herbert J. Klausmeier
Director, R & D Center
Professor of Educational
Psychology

Burton W. Kreitlow
Professor of Educational Policy
Studies and of Agricultural
and Extension Education

Richard G. Morrow
Assistant Professor of
Educational Administration

Wayne Otto
Professor of Curriculum and
Instruction (Reading)

Milton O. Pella
Professor of Curriculum and
Instruction (Science)

Thomas A. Romberg
Assistant Professor of
Mathematics and of
Curriculum and Instruction

Richard L. Venezky
Assistant Professor of English
and of Computer Sciences

MANAGEMENT COUNCIL

***Herbert J. Klausmeier**
Director, R & D Center
Acting Director, Program 1

Mary R. Quilling
Director
Technical Section

Thomas A. Romberg
Director
Programs 2 and 3

James E. Walter
Director
Dissemination Section

Dan G. Woolpert
Director
Operations and Business

* COMMITTEE CHAIRMAN

STATEMENT OF FOCUS

The Wisconsin Research and Development Center for Cognitive Learning focuses on contributing to a better understanding of cognitive learning by children and youth and to the improvement of related educational practices. The strategy for research and development is comprehensive. It includes basic research to generate new knowledge about the conditions and processes of learning and about the processes of instruction, and the subsequent development of research-based instructional materials, many of which are designed for use by teachers and others for use by students. These materials are tested and refined in school settings. Throughout these operations behavioral scientists, curriculum experts, academic scholars, and school people interact, insuring that the results of Center activities are based soundly on knowledge of subject matter and cognitive learning and that they are applied to the improvement of educational practice.

This Technical Report is from Phase 3 of the Project on Prototypic Instructional Systems in Elementary Mathematics in Program 2. General objectives of the Program are to establish rationale and strategy for developing instructional systems, to identify sequences of concepts and cognitive skills, to develop assessment procedures for those concepts and skills, to identify or develop instructional materials associated with the concepts and cognitive skills, and to generate new knowledge about instructional procedures. Contributing to the Program objectives, the Mathematics Project, Phase 1, is developing and testing a televised course in arithmetic for Grades 1-6 which provides not only a complete program of instruction for the pupils but also inservice training for teachers. Phase 2 has a long-term goal of providing an individually guided instructional program in elementary mathematics. Preliminary activities include identifying instructional objectives, student activities, teacher activities, materials, and assessment procedures for integration into a total mathematics curriculum. The third phase focuses on the development of a computer system for managing individually guided instruction in mathematics and on a later extension of the system's applicability.

TABLE OF CONTENTS

Part I

Page

List of Tables viii

List of Illustrations ix

Acknowledgments x

Abstract xiii

I. PROBLEMS OF COMPUTER-FACILITATED MANAGEMENT IN MODERN MATHEMATICS CURRICULA 1-43

- 1.0 General Background and Nature of the Study. . . 1-5
- 1.1 The Systems Approach: Basic Systems Concepts Applicable to Curricula 5-9
- 1.2 Development of A Systems Approach to Curriculum 9-12
- 1.3 Systems Concepts Applied to the Mathematics Curriculum: The State of the Art 12-14
- 1.4 Compatibility of a Systems Approach with Trends in Modern Mathematics Education. 14-25
- 1.5 Problems of Individualizing Instruction in Mathematics 25-31
- 1.6 Problems of Educational Measurement 31-43
  - 1.6. The Need for New Evaluation Metaphors . . 34-43
    - 1.611 Item Difficulty 36-38
    - 1.612 The Assumption of Normality . . . 38-40
    - 1.613 Test Reliability 41-43

II. MANAGEMENT SUPPORT SYSTEMS: THE MEASUREMENT INFORMATION COMPONENT 44-90

- 2.0 Introduction 44-45
- 2.1 Characteristics of Criterion-Tests 45-53
  - 2.11 Absolute vs. Relative Measures 45-46
  - 2.12 Content Validity 46-48
  - 2.13 Parallel Item Selection 48-50
  - 2.14 Minimal Test Length 50-51
  - 2.15 Criterion Selection 51-53

CONTENTS (continued) Page

2.2 A Heuristic for Criterion-Referenced Tests . . 54-60

2.21 Assumptions of the Item-Sampling Model. . 55-57
2.22 Implications for Homogeneous Grouping . . 57-60

2.3 A Formal Theory of Criterion-Tests 60-80

2.31 Functions of A Theory 60-61
2.32 Order of A Model. 61-63
2.33 Basic Definitions for A First Order Theory. 63-64

2.331 Specified Content Objective. . . . 64-66
2.332 Proficiency and True Score 66-69
2.333 Criterion-Referenced Test 69-71

2.34 Sources of Measurement Error. 72-75
2.35 A CRT Performance Model 75-80

2.4 CRT Statistics 80-82
2.5 Implications of the CRT Model 82-87

2.51 For Item Selection. 82-84
2.52 For Test Reliability. 84-87

2.6 Comparison of the CRT Model with other Test Models 87-90

III. MANAGEMENT SUPPORT SYSTEMS: THE DECISION COMPONENT 91-145

3.0 Introduction 91-95

3.01 The Role of Computer and Teacher in Semi-Automated Instructional Systems . . 92-95

3.1 Process vs. Quality Control 95-99
3.2 Specification of Instructional Design Requirements 99-103
3.3 Specification of Acceptance Requirements. . . . 103-108
3.4 Principles of Inspection Sampling of Learning Products 108-122

3.41 Test Outcomes as Points in a Probabilistic Sample Space 111-117
3.42 Scoring Formulas and Data Reduction . . . 117-119
3.43 Lattice Diagrams, CRT Performance Paths, and Criterion Boundaries 119-122

CONTENTS (continued) Page

3.5 Operating Characteristic of a Single Sampling Plan (SSP) 122-131

3.51 Ideal Sampling Plan Characteristics . . . 123-125
3.52 Admissable Sampling Plan Characteristics 126-127
3.53 Impirical Considerations for Selecting An SSP 127-130
3.54 Procedures for Selecting An SSP 130-131

3.6 Curtailing the Single Sampling Plan 131-135
3.7 Decision-theoretic Considerations in Sampling Plan Selection 135-138
3.8 The Sequential Probability Ratio Test (SPRT) 138-141
3.9 Summary Comparison of Curtailed and Single Sampling 141-145

IV. IMPLICATIONS FOR CONTINUED PROGRAMMATIC RESEARCH . . 146-177

4.0 Overview 146-147
4.1 Implications for Further Development of Criterion-Reference Test Theory 147-165

4.11 Affective Validation 149-154
4.12 Functional Validation 154-163
4.13 Formal Validation 163-165

4.2 Implications for Curriculum and Instruction . . 166-175

4.21 Identifying Curriculum Hierarchies . . . 166-167
4.22 Predictive Learning Theory 167-173
4.23 Implications for Assessing Instructional Effectiveness 173-175

4.3 Implications for the Systems Approach to Education 175-176
4.4 Implications for Teacher Education 176-177

Appendix A Sample Criterion Tests and Test Results 178-186
Appendix B Sample Computer-Generated Criterion Tests 187-210
Appendix C List of Symbols and Notation Used 211-213
Appendix D Computer Programs for Generating OC and ASN Curves 214-219
Appendix E Notes on the Sequential Probability Ratio Test 220-227
Appendix F Prototypic Program and Data Files 228-262
Appendix G Figures 1-16 263-277

References 278-285

LIST OF TABLES

TABLE PAGE

3-1 Specification of Design Requirements 100

3-2 Illustrative Design Requirements 111

3-3 Illustrative Acceptance Requirements 112

3-4 Probabilities Associated with Response Patterns and Test Scores. 115

3-5 Comparison Data on Acceptance Requirements . . . 129

LIST OF ILLUSTRATIONS

FIGURE PAGE

1 A General Model for the ESM Curriculum 264

2 A Pattern for Instruction 265

3 Lattice Diagram (n=3) 266

4 Lattice Diagram (n=20) 267

5 OC Curves for Fixed N and Selected Criteria (n=5). . . 268

6 OC Curves for Fixed N and Selected Criteria (n=20) . . 269

7 OC Curves for Fixed N and Selected Criteria (n=25) . . 270

8 Step Function OC 271

9 OC with Region of Indifference 271

10 Decision Outcome Probabilities 271

11 ASN Curves for Fixed N and Selected Criteria (n=5) . . 272

12 ASN Curves for Fixed N and Selected Criteria (n=20). . 273

13 ASN Curves for Fixed N and Selected Criteria (n=25). . 274

14 Comparative Sample Sizes for Various Sampling Plans. . 275

15 Random Sample of 2-Digit Multiplication Problems Generated by MATH/CMS 276

16 Hypothetical Proficiency Development Curve 277

ACKNOWLEDGMENTS

The task of properly acknowledging all who contributed to this work presents to the author perhaps the most difficult challenge of the entire effort. This is partly so because, in the case of Vere DeVault--my major professor, inspirer, and guide--no expression of thanks would seem fully adequate. In the case of other committee members, particularly Frank Baker and Anne Cleary, their persistent and perceptive probing led not only to a better final product but also, along the way, to some energetic exchanges concerning our differences in regard to the sum and substance of educational measurement. An expression of gratitude to both is in order and hereby tendered, but there lingers the feeling that possibly *touché* would be more appropriate than *merci*.

The contributions of others have been made in perhaps less dramatic and certainly less traumatic form. J. Fred Weaver, Al Yee, and Dan Anderson have, over the many months of this investigation, contributed in a variety of ways for which I wish to express my appreciation. Harry Silberman and John Coulson of Systems Development Corporation provided the author with a unique opportunity to work on the criterion-test theory. To James Bavry of the Wisconsin Research and Development Center goes my thanks for working out the computer programming needed to graph the operating characteristics

and average sample number associated with the sequential probability ratio test. Others at the R and D Center who were particularly helpful, especially in regard to measurement questions, were Tom Romberg and Tom Fischbach.

My appreciation is also expressed for the excellent work done by the artists, publication and technical personnel, and secretaries of the Center all of whom met the challenge of the Greek alphabet and my penmanship with unfailing good grace and impressive competence. Special mention is merited by the fine work performed by our project secretary, Ethel Koshalek; the thesis typists, Jan Sluga, Dorothy Egener, and Darlene Barkema; and to Doris Ardelt and Julie Erjavec who together managed the final production of the paper.

Use of the University of Wisconsin Computing Center was made possible through support, in part, from the National Science Foundation, other United States Government agencies and the Wisconsin Alumni Research Foundation (WARF) through the University of Wisconsin Research Committee.

This work was made possible through the facilities of the Wisconsin Research and Development Center for Cognitive Learning, Herbert J. Klausmeier, Director.

ABSTRACT

Individualization of instruction and the development of inquiry learning techniques are two major areas of interest in mathematics education today. Those teachers who would implement individualized programs of instruction and would encourage pupils to take increasing responsibility for inquiring about mathematics objectives of their own choosing face many problems of classroom management, data acquisition, record-keeping, diagnosing, prescribing, and decision-making for which traditional classroom management practices are poorly suited. This dissertation treats several problems of instructional management encountered in situations which emphasize self-selection and self-pacing principles. Primarily, these problems deal with the efficient utilization and allocation of available human and material resources to create an operational, individualized, inquiry-learning environment.

The systems approach to education is first explored to determine the extent to which systems disciplines, such as utility theory and operations analysis, can be applied to facilitate classroom management and decision-making procedures. This exploration leads to an examination of test-design, since tests comprise the basic means of obtaining decision-data. Traditional, norm-referenced test theory is shown to suffer a number of disadvantages from the instructional management viewpoint. Therefore, an approach toward test design is developed based on certain principles of item-sampling and criterion-testing.

The criterion-test is seen as a sequence of independent Bernoulli trials on items randomly selected from a well-defined population of items. The item pool is operationally defined by a Specified Content Objective or, equivalently, by a set of item-generation rules. This approach leads to a binomial test model that is utilized in the design of acceptance sampling procedures. These procedures minimize testing time required to obtain suitably reliable pupil proficiency profiles.

A prototypic computer-managed system of mathematics instruction demonstrates the use of criterion-referenced test theory and acceptance sampling principles. The prototype is restricted for demonstration purposes to the universe of integers, numeration systems two through ten, simple equivalence statements, and the four fundamental arithmetic operations. Data gathering, decision-making, and information handling techniques are also illustrated by the program and the protocols involved in using its associated data files.

Suggestions are offered for reducing the testing time needed to detect mastery attainment levels consistent with Neyman-Pearson theory. The relationship between mastery criteria and such sampling plans as single sampling, simple curtailed testing, and the use of the sequential probability ratio test is discussed. Applications are indicated in the area of computer-generation of test items and automated administration of criterion-referenced tests of mastery in selected arithmetic skills.

The use of criterion-referenced test results, evaluated by sequential analytic techniques indicates promise for reducing testing time and costs for specified behavioral objectives. This, in turn, promises the possibility of designing improved and extended capabilities for computer-assisted instructional management systems.

CHAPTER I

PROBLEMS OF COMPUTER-FACILITATED MANAGEMENT IN MODERN MATHEMATICS CURRICULA

1.0 General Background and Nature of The Study

One of the more difficult problems encountered in the design of computer-facilitated systems of instructional management is called the decision problem. Basically, it is a problem involving the assessment of student performance in such a way that one can select from among available alternatives those learning activities which would be most appropriate to undertake next. An essential part of the art of teaching, this "diagnosis and prescription" function has in recent years become more difficult for teachers to perform as innovations stressing individualized instruction have been implemented. As traditional group instruction practices evolve more and more toward arrangements permitting continuous pupil progress, traditional instructional management techniques devised in the 19th Century for handling lock-step classes have become less and less effective. It is the purpose of this dissertation to develop a decision system based on a relevant theory of criterion-referenced tests and explore its potential for solving some problems encountered in the management of individualized mathematics curricula.

The alternatives for designing a decision system appear at first glance to be numerous. One might proceed, for example, to set up a mathematical learning model and employ probability theory, decision theory, utility theory, or Bayesian inference techniques to optimize some set of parameters which describe the learning process. This general kind of approach has been explored rather extensively (Bush and Mosteller, 1955; Davidson and Suppes, 1957; Estes and Suppes, 1959; Karush and Dear, 1964; Carne, 1965; Fogel _et al_, 1966; Groen and Atkinson, 1966; Smallwood, 1967). Due to a variety of well known problems, such an approach remains one of theoretical interest. Basically, the models are not sufficiently flexible or realistic to handle day-to-day problems of instructional management in the classroom (Lord and Novick, 1968, p. 2).

Another option lies via the route that might be generically described as pattern recognition (Feigenbaum and Feldman, 1963). Here the idea is to use computer information processing models to identify, by induction (Hunt _et al_, 1966; Press and Rogers, 1967), factor analysis (Forgy, 1965), or other means, those attributes which assist in the diagnosis of learning needs. Flanagan (1967), for example, describes how this approach might be used in connection with project PLAN to associate individual pupil learning needs with appropriate instructional packets. Project ULTRA (Spector, 1965) is another example of developing a large scale combined management and instructional program which anticipates the use of computers to analyse large data bases to identify learning patterns which need attention. Although this is a promising approach, these large scale systems tend, at present, to make excessive demands on computer storage and computing time and therefore face serious obstacles from an economic standpoint (Kopstein and Seidel, 1968; Oettinger, 1968).

Some projects emphasize the use of computers to store large numbers of questions in a "curriculum data base" and then administer a sequence of drill and tests utilizing decision rules built from empirical evidence gained in pilot studies. Suppes' (1968) Drill and Practice Program is of this type. Arithmetic problems are pre-categorized into content blocks and sorted by difficulty level. Decisions regarding level of drill, need for review, and testing are made on the basis of probability distributions built from previously gathered empirical data for similar groups of students. Other approaches (Coulson *et al*,1968) attempt to reduce costs by using the computer in a batch rather than interactive mode, primarily for data analysis and report writing but not for either testing or instruction.

These various approaches to instructional management may be conveniently classified in two broad categories. The first is based on mathematical learning models, the second on statistical models of behavior or processess. The system techniques described in this paper belong philosophically to the second category but differ from most work done previously in the kind of test instrument used to generate the data and to some extent in the role played by the computer. Rather than use tests built according to the principles of classical test theory (Gulliksen, 1950; Lord and Novick, 1968), a strict item-sampling model for criterion-referenced tests will be described which promises a number of advantages for applications to instructional management system.

Classical test theory suggests how to build instruments well suited for relative ranking, grading, and selection operations. However, following a line of thinking begun by Carroll (1962, 1963) and elaborated by Bloom (1968), we find it appropriate to consider a procedure that may be more useful in system design; namely, measuring absolute levels of behavioral proficiency through the use of a particular kind of criterion-referenced test. The potential advantages of this approach over one based upon classical norm-referenced test theory include the generation of tests possessing a high degree of content-validity, simplified record keeping, minimal data storage requirements, efficient implementation of the computer's capability for randomly generating test items, better diagnosis of learning difficulties, and improved prescriptions for these difficulties. The role of the computer in this system is roughly comparable to that of an industrial quality control inspector. To a large extent, the computer is used to get the teacher out of time-consuming administrative and clerical duties and to bring him more directly into an individualized instructional role.

Thus, the purpose of this study is to design a prototypic instructional management system that capitalizes on the potential advantages listed above. Specific problems treated include the theoretical development of an item-sampling model of test construction and its application to a system designed for management of a continuous progress environment of mathematics learning at the intermediate grade level. The model will suggest techniques for minimizing costs and time required for testing as well as methods for generating quality-assured tests in an economical fashion. A number of important implications for further study through the use of the prototypic management system as a research vehicle will become evident particularly with regard to (1) the design of experiments for testing hypotheses concerning the relative effectiveness of alternative instructional treatments and (2) task analysis and the identification of hierarchical curriculum structures.

1.1 The Systems Approach: Basic Systems Concepts Applicable to Curricula

A system may be defined as an organization of parts into an interrelated operational unit. The parts of the system are called components. A real world system exists for the sake of its output. The output therefore is one of the important factors by which the purpose of the system can be described and evaluated. The system is constructed and adjusted to maximize specified kinds of output.

In order to achieve desired system performance, the necessary inputs must be made available to the system. Every complex system operates under certain constraints such as cost or time limitations. In order to be efficient, the costs of system operation must be minimized and the profits or utility maximized.

In complex systems, there are numerous decisions to make between conflicting demands and priorities within the system. In such situations, it is typical to give subjective values to various kinds of system outputs as measures of system utility. Utility is found to vary as the adjustable features of the components, called system parameters, are made to vary.

The general systems problem can be formulated in terms of finding those parameter values which maximize utility and minimize cost. Fundamentally, one can say that systems analysis is designed to enable decision-making which optimizes system performance on the basis of a utility index.

Systems analysis has value in situations which are too complex to be guided effectively by the independent management of the component parts. Systems analysis is essential where conflicting demands exist within the system. One purpose of the analysis is to find techniques for adjudicating the conflict in such a way as to optimize performance.

Complex systems are often studied effectively through the use of a model. The model includes provision for system inputs as well as provision for the adjustment and variation of system parameters in order to observe the idealized performance of both the total system and its subsystems. Since the model is a simplified version of the real world, the performance of the model will not perfectly predict system performance. However the sacrifice in terms of accuracy of prediction is compensated by the manageability of the model. The model is therefore useful in providing insight into the more complicated functioning of the real world system (Cogswell, et al, 1964).

There are a number of other important uses that can be made of models. Often one must prepare people (managers) to handle situations that are too expensive or difficult or perhaps too dangerous to attempt to practice learning the needed management skills in the real world situation. In such cases, a model can be used to simulate the situation and the system manager can make decisions and observe the consequences as determined by the functioning of the model. One form of such a simulation exercise is known as gaming (Eberhart, 1966; Flexman and Horowitz, 1966). A related form of simulation is used to train people to perform complex tasks, such as flying aircraft or making a moonlanding (Stein, 1967; v. Braun, 1966). Corresponding efforts have been extended to include preparation of teachers in simulated classroom environments in recent years (Cruickshank, 1966; Dettre, 1967; Gerlach, 1967).

Because the model provides a method of organizing ideas to predict system performance, it serves a function similar to that of theory in the so-called "hard" sciences. That is, the model serves to provide explanations of observed phenomena as well as predict future performance given the conditions of system operation. Thus one often finds the terms model and theory used interchangeably.

The systems approach to education ordinarily must be carried out in several steps. First of all there is the need to specify the problem, task, or situation which a system will be designed to handle. Ideally, a number of system models might then be hypothesized to meet the requirements of the situation. From among the competing models, one model may be selected as a candidate and its performance studied. However, before the model performance can be examined the system parameters must be identified and their values estimated. The model can then be tested and evaluated to see if it satisfies the requirements, i.e., if it helps solve the problem or possesses predictive validity, or enables improved management of a complex situation.

As indicated earlier, optimization of system performance is the general goal of systems analysis. However, the term optimization encompasses many specific interests one may have in studying the system. System efficiency is often an important consideration, system reliability is another, and system stability is yet another. As will be developed later, each of these has special relevance to the problems encountered in mathematics curriculum construction. System stability often requires that inputs and parameter settings be modified by a knowledge of the output. This stabilizing technique is referred to as the use of feedback.

The development of a system model for some real world application is largely an art rather than a science at the present time. It requires creativity, insight, and experience on the part of the modeler. However, there are some general guidelines to follow. Identifying and defining all the essential system components is a first concern. Once the system components are selected, the system parameters must be defined. Next, parameter values are estimated if data is available from previous research.

If such data are not available, parameter values may be arbitrarily assigned. In the latter case, the parameters are not treated as descriptors of an existing real world system but rather as control variables which may be used to predict the performance of an isomorphic real world system. Often a simulation via computer may then be carried out to study properties of the model or to enable the modeler a means by which he can gain more finely detailed understanding of a complex process.

Once a valid model is constructed, i.e., a model which helps to solve some practical problem, one can consider optimization procedures. Finding the parameter values which optimize system performance is part of what is known as the decision problem. An elaborate set of procedures for decision making has grown out of the classical axioms of decision theory formulated by Morgenstern and von Neumann (1944). Methods of decision analysis relevant to learning models have been explored in connection with these axioms as indicated previously (p.1).

1.2 Development of A Systems Approach to Curriculum

Since curriculum terminology often lacks specificity, some fundamental definitions may be helpful at the outset. Definitions of curriculum cover a broad spectrum. Among the variety of definitions cited by Goodlad (1960), one notes a common point of agreement; namely, that a curriculum is a plan, a plan for instruction. A theory of curriculum, then, may be described as a model for planning instruction. The theory, in this view, is a stable framework about which one builds particular instructional plans. The purpose of the theory is to reduce the complexity of decision-making, to promote consistency of instructional practice, and to improve quality of learning in each individual pupil.

In general, the need for theory arises in situations which are too complex to understand, to explain, to predict, or to manage without the assistance of some simplifying device. Simplification is achieved by taking risks. Certain complicating features are disregarded. Only the most essential characteristics required for explanatory, predictive, control, or management purposes are retained in the model.

Instruction, in general, and mathematics instruction in particular is the interactive process that results from the implementation of a curriculum. The process of instruction involves two distinct subprocesses, learning and teaching.

Learning is a process that has an enabling character; it provides the learner with a new capacity for acting, a potential for thinking and behaving that, prior to the learning experience, did not exist in him. Teaching is a process which by preceptive, didactic, or dialectic techniques attempts to facilitate and guide learning toward goals specified in the curriculum. The intended product of instruction is the generation of a new potential for specified overt or covert behavior in the learner.

It should be noted that learning is not simply behavior nor a change in behavior. It is rather a process that results in a capacity for new behavior. Inferences concerning the state of this learned potential are made by evaluating the kinetic behavioral displays elicited by suitable test stimuli. Nevertheless it is the unobservable potential for specified behavior, rather than the behavior itself, that must be developed, maintained, and extended as the primary objective of instruction.

Implicit in this delineation of plan (curriculum), process (instruction, learning), and products (capacity or potential for specified behavior) is the inescapable responsibility of the curriculum designer to set goals for instruction in the context of a value system. The consideration of values often brings him into a conflict situation. For example, individualized attention to learning needs may be a highly valued characteristic to be incorporated in the instructional design. On the other hand, dollar cost and teacher time must be considered as valued dimensions of instructional design also. Maximizing one value minimizes another and a conflict exists.

The problem of selecting design specifications which optimize subjective or objective values is one for which certain techniques of system analysis were first developed in areas of endeavor outside education. In recent years, however, educators have been giving increasing attention to the systems approach in education.

The application of the systems approach to curriculum construction is best described as being in its infancy. Little is known concerning the extent of applicability or the effectiveness that system management techniques, probabilistic decision theory, or value theory may have in curriculum construction. Nevertheless, there are a number of persistent unsolved problems which have arisen in programs of research in individualized mathematics instruction that may yield to a system approach. Specifically, these are (1) providing the means for teaching intermediate grade children how to make individual inquiry into a self-selected area of mathematical knowledge; (2) developing and managing methods of bringing human and material resources into effective contact with inquiring learners; (3) identification of the decision-points, alternatives, and values encountered in such a system that can be effectively analyzed through computer assistance; and (4) the applicability of probabilistic and deterministic decision algorithms to the management and assessment of the learning process which occurs in a continuous progress environment.

These are difficult and significant curriculum problems. The manner in which one proceeds toward workable solutions is in the spirit of bootstrapping. A start toward the development of a systems model applicable to the mathematics curriculum has been suggested recently (Romberg and DeVault, 1967). The decision system treated in this dissertation should contribute to this development of a systems approach to mathematics education in the sense of devising general techniques for maximizing expected utility by applying a systematic decision-making process to the consideration of values contingent upon available alternatives for reaching specified objectives.

1.3 Systems Concepts Applied to the Mathematics Curriculum: The State of The Art

As in other areas of education, the application of systems concepts to the mathematics curriculum is of recent origin. Traditional approaches to research in mathematics curriculum are described by DeVault (1966) as falling into descriptive, relational, and experimental categories. Much of this research has had little effect on classroom practice however. Explanations for this phenomenon range from such general criticism as that voiced by Campbell and Stanley (1963) concerning the effects of faulty experimental design to arguments such as Armstrong's (1966), suggesting a "failure to consider all of the major input elements."

One stream of activity designed to improve the effectiveness of curriculum research can be identified in terms of a trend toward the application of systems analysis. Macdonald (1966), for instance, suggests a curriculum system having four components which he calls curriculum, learning, instruction, and teacher. Armstrong (1966) derived four similar components for a later version of the model called curriculum, learner, instruction, and teacher. This particular trend toward identifying the components of curriculum has appeared in the mathematics curriculum literature most recently in connection with Romberg and DeVault's model. Here the components are described as the content, learner, teacher, and instruction. The relation of the components are suggested by the diagram shown in Figure 1, Appendix G.

Beyond this point, there is little in the way of organized theoretical development to report. This marks the starting point, in effect, from which the further development of a curriculum system begins.

An outline of programatic research toward which the results of this dissertation would contribute is as follows. Given the components of the Romberg-DeVault model of mathematics curriculum, it would be necessary to identify system parameters as the next step. Once that is accomplished, it may be possible to construct a management model of a classroom situation which can be used for some of the purposes proposed earlier. Specific possible uses for such a model would include the study of system reliability, efficiency, and stability in the management of a continuous progress environment designed to enable intermediate-grade pupils to attain mastery of designated portions of mathematics. From such a study, there could come certain practical applications particularly in the application of stochastic methods to instructional management and in the computer-generation of evaluation instruments. Such techniques, if successful, should provide an adequate basis for utilizing computer assistance in securing admissable, if not optimal, performance from a system of individualized mathematics instruction.

1.4 Compatibility of a Systems Approach with Trends in Modern Mathematics Education

The elementary and secondary school mathematics curricula may be conveniently examined in three stages of development. Divisions between the three stages are not sharply marked in terms of time. However, the approximate times delimiting the major effective points at which new directions were taken are the decades following World War I and World War II respectively. (Eby, 1952; Cremins, 1961;

DeVault and Kriewall, 1969).

The first era, dating from the beginnings of the American School System, established strong traditions that are still evident in today's mathematics instruction. The classic textbooks of this early time were characterized by their brevity and topical inclusiveness. All treated the four fundamental operations on whole numbers and fractions, units of measure and conversion problems, and monetary topics. The books were typically small, uninteresting by modern standards, intended for use by literate people of any age, and designed to develop skills that would be useful in the personal and commercial aspects of adult life.

Following the Civil War, as the forces of industrial revolution and immigration made their impact, schooling experienced major changes. Age-grouping into grades became necessary for effective administration, elementary and secondary schools were established, universal and compulsory education became accepted in one State after another, and great stress was laid on discipline as the pupil-teacher ratio mushroomed.

The rationale which gave arithmetic its place in the colonial and antebellum curriculum was its practicality. As American secondary schools developed toward the end of the 19th Century, the need and purpose for algebra and geometry in the curriculum was defended largely on the basis of faculty psychology. It was argued that the discipline offered by such study for the faculties of reason, memory, and neatness could be expected to produce more logical adults having good habits of neatness and capable of remembering instructions clearly. Similar arguments gradually were added to those of utility in defense of elementary school arithmetic also.

One effect of this view that should be especially noted here is that a kind of mastery learning became the standard by which instruction was judged. Instruction which produced children capable of speed and accuracy in computation and recollection of rules and facts was considered the mark of a good teacher. Those who failed to develop the children's talents in regard to these criteria were judged poor teachers.

According to the accounts of Rice (1893), emphasis on mastery and discipline had combined with a dead formalism in all content areas which resulted in increasingly severe measures ironically intended to overcome the stultifying effects of the methodology. Arithmetic emerged as one target for reform with many calling for both its enrichment and abridgment (Committee of Ten Report to the National Education Association (NEA), 1893). At the secondary level, similar concerns for a more humane and more relevant mathematics curriculum were voiced by Perry in his address to the British Royal Society (1901) and by E. H. Moore in an address to the American Mathematical Society (1902).

The decade from about 1892-1902 marked the beginning of a phase in which formalism, discipline, mastery, subject-centeredness, and similar related ideas were severely attacked by educational reformers. By the beginning of the Interbellum Period, such attitudes toward schooling would be eclipsed by concern for the child's interests and needs, concerns which grew out of new philosophies of education by James and Dewey as well as from new views of mind introduced by G. S. Hall, Freud, Thorndike, and others.

Elementary school arithmetic, in the second major era of its development, was systematically reorganized according to what were considered at the time to be scientific tenets of Social Utility Theory (SUT), (Monroe, 1917; Wilson, 1949) and Stimulus-Response psychology (Thorndike, 1922). Grade-placement of topics were recommended on the basis of research which associated each topic with a parameter called mental age (Washburne, 1931). Serious efforts were made to apply controlled research to discover optimal algorithms for attaining speed and accuracy in computation (Brownell, et al, 1948). Efforts were also made to make arithmetic significant in terms of its social applications (Brueckner and Grossnickle, 1947; Tenth Yearbook of the National Council of Teachers of Mathematics (NCTM), 1935).

The influence of faculty psychology steadily waned following the classic experiment of Thorndike and Woodworth (1901) which showed that the expected transfer of good habits associated with the general discipline of certain faculties did not occur. Accordingly, the importance of arithmetic in the elementary school curriculum declined and the practice of incidental instruction increased, especially at the primary and early intermediate levels.

The secondary mathematic curriculum underwent a slow parallel change during the Interbellum era also. The urging of Moore and Perry to make practical usefulness the guide to selection of topics resulted in the introduction of analytic geometry in the form of graphing as a new topic. Concurrently, there was a gradual increase in emphasis on verbal problems selected from areas of science. As the result of a study sponsored by the Mathematical Association of American (MAA), a report was published in 1923 that urged new Junior High courses which would break down distinctions between algebra, geometry, and arithmetic. The report reflected the influence of SUT by advising greater emphasis on preparation for everyday life with new topics to be selected which dealt with savings, investments, insurance, taxes, etc.

By 1940, concern for child development competed strongly with SUT as a basis for curriculum construction. The report of the Progressive Education Society (PEA, 1940) stressed the child's needs which arise in persistent life situations. World War II raised to a peak the concern for democratic ideals in schooling. During this period in which "child-centered", "democratic," and "socially meaningful" were terms which denoted the growing key concerns, mathematics found a new role in the curriculum as the means by which the "free play of intelligence" could be cultivated, principally by virtue of its ability to enhance problem solving capabilities.

In these and subsequent years, the discussion of curriculum became more interdisciplinary in nature, the project method grew in popularity, and mastery of fundamentals began to be looked upon as an evil, associated with past excesses of rote learning, meaningless verbalization, and deadening drill.

The influence of professional mathematicians was confined to the secondary school level during the Interbellum period. A report of the Joint Commission on the Place of Mathematics in Secondary Schools sponsored by the MAA and NCTM (15th Yearbook, NCTM) identified goals that stressed clear thinking, strong skills, and healthy attitudes toward mathematics, approximately in that order of importance. By way of contrast, the PEA (1940) report urged that such objectives as personal living, personal-social relationships, social-civic relationships, and economic-career relationships be given first consideration in elementary school curriculum development.

The end of the second era in mathematics curriculum development saw elementary and secondary objectives sharply divided. The elementary school was dominated by concerns for the child and society while the secondary school remained more strongly oriented toward a disciplinary approach in spite of weak attempts to liberalize the subject matter. Controversy between the elementary and secondary levels sharpened into an unhappy division between what appeared to be camps devoted to a child-centered and subject-centered curriculum, respectively.

The Postwar era of mathematics curriculum development has been characterized by the influence of the academicians in K-12 curriculum construction on the one hand, and by the behavioral psychologists on the other.

A wide spectrum of major issues in mathematics curriculum construction arose following World War II covering teacher training;

special provisions for both slow and fast learners; intellectualization of the curriculum; use of television, teaching machines, programed instruction, and computers; development and application of multi-media materials; and the preparation of more rigorous mathematical textbooks and related materials. Rapid advances in technology along with a growing respect in the public eye for scientific and technological endeavors helped swing the pendulum toward emphasis on academic concerns.

The secondary school was the first to feel the main effects of what was named the modern mathematics revolution. In 1952, the University of Illinois Committee on School Mathematics (UICSM) was formed to investigate and find ways of improving what appeared to be declining standards of mathematical performance in incoming freshmen. There followed in rapid succession the formation of the Commission on Mathematics in 1955 and of the School Study Mathematics Group (SMSG) in 1957 to implement its recommendations. Numerous groups flourished suddenly in the period from 1957 to about 1961 as money became available to rebuild the curriculum ostensibly for the sake of national defense. By the early Sixties, UICSM, SMSG, and the Madison Project (Davis, 1965) turned their attention to the elementary school mathematics curriculum.

It is not necessary for the immediate purposes here to document fully this vast activity. The essential point is that in the rush to improve the mathematics curriculum, there has been perhaps too much emphasis on one possible method for seeking improvement. That method, as many mathematicians see it, is to build first a familiarity with mathematically important notions such as the structure of the real number system and basic ideas of synthetic and analytic geometry (CCSM, 1963, p.8).

One of the essential problems often ignored by this approach is that parents demand the development of functional and foundational competence on the basis of a value system different from that of the academician. What is clearly needed is the option to select locally valued alternatives and build a sound mathematical program in keeping with it.

It is therefore essential to consider curriculum systems capable of adjusting to the relative values placed on functional, foundational, and formal mathematics as well as to the available resources for handling instruction in each area. This is a basic purpose of the systems approach to curriculum and the projected goal of utilizing some automation in instructional management.

It is not likely that all cognitive levels of mathematics learning can be effectively improved simply by utilizing better management techniques. This is why we distinguish between three historically significant levels of instruction. The strategy is designed to more efficiently handle the first two levels so that the professional competence of the teacher can be brought to bear more effectively at the higher cognitive learning and evaluation levels.

Functional mathematics is taken to mean the kind of knowledge and skill which assists a person to function in adult life without embarrassment. The ability to perform basic arithmetic operations on whole numbers and fractions is essential in this regard. In spite of the poor image given to this part of mathematics, there is general agreement that the development of such skill is important for all students.

Foundational mathematics includes the skills and concepts needed for successful study in related areas of inquiry such as the physical sciences, psychology, economics, and so on. This presumably includes functional mathematics as a subset. However, there are many areas in the traditional domains of algebra and geometry which have foundational value but do not possess obvious functional value in ordinary adult life situations.

Finally, the new mathematics has stressed the unifying ideas of the formal structures of mathematics. These have been proposed for inclusion in the curriculum for a number of purposes. It has been asserted, for instance, that emphasis on certain formal aspects of mathematics will promote clearer communication of mathematical ideas and therefore more efficient progress and mathematical growth by children. It has been said that what is called functional and foundational mathematics here is not really mathematics at all and without the insemination of certain formal notions, the child is left a mathematically illiterate adult.

The three labels help sort out purposes in today's conglomerate of mathematics curricula. Mathematicians and to some extent, mathematics educators have stressed over the past decade the importance of formal or "clean" mathematics, suitably modified to the developmental needs of children, to be sure. However, a basic contention appears not to be borne out in classroom practice. And that contention is that by beginning with the suitably adapted notions of formal mathematics, the child will learn rather incidentally and painlessly the necessary functional skills and concepts (CCSM, 1963). Little hard evidence is yet available to support or refute this, nevertheless a number of informal sources of evidence suggests that the belief is not entirely justified. It is not difficult to find, for instance, students in the Junior High level awkwardly performing long division by the ladder method in the cultivated but mistaken belief that this is one of the ultimate techniques advocated in the new mathematics. Nor is it rare to find pupils who, contrary to predictions, fail to outgrow the tables of basic facts that are used to look up rather than memorize results of basic operations. Romberg (1968) reports failures correctly to reduce fractions to lowest terms occurred with twice the frequency in students in modern courses as compared with those in traditional courses. From these and other similar reports, the computational proficiency of elementary and secondary school children appears to be remaining at unsatisfactory levels in many cases with some evidence suggesting declines in performance associated with some of the new programs. Except for cases such as the Madison Project where elementary mathematics is taught by Ph.D. level mathematicians, it appears that emphasis on "structures" in mathematics more often than not at the present time tends to produce in the typical classroom only a new formalism rather than mathematical insight and that, as a consequence, functional aspects of mathematics are ineffectively learned and certainly not mastered. It further appears that one remedy may lie in the direction of a well managed and balanced emphasis on mastery learning of functional mathematics as a preparation for later development of "mathematical literacy." Appropriate emphasis on functional skills should also help provide the necessary background for the efficient development of foundational skills and concepts as needed by individual students.

This position specifically rejects, however, a return to the old emphasis on rule and rote. Nor does it say that functional mathematics is the whole of mathematics or its essence. Similarly rejected is a position which advocates the instruction of functional mathematics solely because of its utilitarian value. In terms of Macdonald's (1964) description of curriculum, functional mathematics is regarded here as having both consequential and existential importance, i.e., the ability to function effectively in situations demanding fluency is an important consequential goal of education and the evident practicality of such ability to the child in his immediate life is an important existential goal.

The most appropriate approach to take in further developing the mathematics curriculum at this point in time, it is hypothesized here, is one which keeps the many considerations inherent in curriculum construction well balanced. Thus it is appropriate to look to the systems approach for techniques that optimize the performance of a complex system of conflicting demands and utilities inherent in a curriculum. However, the impact of system analysis on the curriculum in general has been marginal and in the case of the mathematics curriculum there are only the few instances cited earlier (p.12) to report. Certainly DeVault's (1966) suggestion that mathematics curriculum research of the future mount daring frontal attacks on the multi-facted problems points to a route which will of necessity require that ways be found to apply the more sophisticated tools of systems analysis if the traverse is likely to be in any sense successful.

1.5 Problems of Individualizing Instruction in Mathematics

Mathematics instruction for students in grades K-12 can be seen in historical perspective as having three component parts. These three components relate to the development of the child's competence in (1) functional skills, (2) foundational concepts and skills, and (3) formal mathematical ability. The first two categories have traditionally occupied the larger share of time in the school curriculum. In spite of continuing efforts to modernize mathematics instruction in such a way as to place greater emphasis on the structural or formal aspects of mathematics, the basic instruction in the schools has not radically changed.

The first two categories of mathematics instruction typically seek mastery learning as the desired outcome, in the sense that Bloom (1968) uses the term. The fundamental task of instructional management facing the teacher in a mastery learning situation can be described in terms of a binary decision problem: given an individual student and a specified behavioral objective, does the student's behavior in the presence of appropriate stimuli indicate the attainment of (appropriately defined) mastery of the objective, or does his behavior indicate that mastery has not been attained? Subsequent decisions should follow from this first decision: If mastery is not attained, what diagnosis can be made of learning deficiences? What does the diagnosis indicate in terms of needed prescriptive remedy? If mastery is attained by some individuals but not a sufficient proportion of the entire class, what individual or group learning experiences should be prescribed?

This fundamental decision, diagnosis, and prescription problem is faced daily by many grade K-12 teachers of mathematics. The solution to the problem is gained in the classroom usually by a combination of experienced judgment, intuition, and guess. However, the problem lends itself to analysis and therefore there is reason to hope that better management of a variety of classroom situations can be achieved.

The approach toward better classroom management practices envisioned here includes some automation through the use of modern computer technology. However, it would be incorrect to think of this as implying the need for some sort of dehumanized, machine-oriented system for developing animal-like mathematical conditioning of children. Rather, this trend is appropriately viewed as a natural extension of previously documented curriculum trends toward a more humanistic and child-oriented curriculum through the use of better classroom management techniques. The key consideration lies in the individualization of instruction.

In general, it can be said that present trends toward the development and use of systems techniques in mathematics education are being prompted by increasing interest in the individualization of instruction and the application of the new technology offered by modern computers to solve persistent problems which arise in this connection. Existing research projects bear certain similarities to one another and to the work proposed here, as well as certain distinctions. Differences are found mainly in the following categories: (1) Use of resources. Material resources used by some projects are restricted to only those generally available in the ordinary classroom. Other projects, e.g. the Oakleaf program of Individually Prescribed Instruction and Project PLAN, rely almost entirely on specially prepared materials. The use of human resources in the various projects varies considerably also. System Development Corporation has a project underway which emphasizes the development of a tutorial community involving parents, teachers, administrators, paraprofessionals, siblings, and peers. At the other extreme, there have been some experiments in computer assisted instruction and programed learning in which human intervention is held to a minimum.

(2) Subject matter and grade level. Research programs vary considerably with respect to these indices. The instructional management system (IMS) now being developed for the Southwest Regional Laboratory (SWRL) is presently focused on grade 1 and mainly devoted to developing reading skills. Another SWRL project involving "unconditionally successful instruction" (USI) uses seventh grade arithmetic as a research vehicle. The drill and practice programs developed by Suppes (Suppes, 1968; Radio

Corporation of America, 1967) assume some form of prior instruction which develops skills in the first instance. Maintenance of skills is emphasized in these projects which cover grades two through six and currently emphasize arithmetic skills and certain reading skills.

(3) Technology. Finally, programs differ in their utilization of computers and especially in the kinds of terminals used to facilitate communication between teachers, pupils, and the materials.

Computer-assisted instruction (CAI) has existed in various forms at the college level in several subject areas for a number of years now. Among the better known are perhaps Illinois' Project PLATO and New York Institute of Technology's Project ULTRA. PLATO employs interactive terminals whereas ULTRA relies primarily on passive terminals such as headsets and audiotapes.

Most systems use computer technology in a command mode. That is, decisions are made on the basis of given rules and the student is expected to conform to whatever requirements are set up by the system. Recent instructional research has begun to make provision for continuous progress based on principles of self-selection and self-pacing, however. Systems employing these principles might use the computer in what could be called a demand mode: the human participants make the major decisions with the assistance of management data supplied upon demand.

Perhaps the one objective these several major efforts have in common is the desire to reach more effectively the individual learner at his level, to individualize instruction so that it is maximally useful, meaningful, and rewarding to the learner. The attempt to individualize instruction has barely gotten underway, however. A number of basic problems persist that prevent the dream from becoming a reality at the present time. One problem frequently cited is the cost of system management hardware. Nevertheless, the use of computers to serve many functions such as record keeping, testing, grading, and instruction is essential if individualized instruction is ever likely to be realized. The cost at present of using interactive terminals for instruction is many times that of conventional grouped instruction. Thus, monetary factors remain an important consideration for the curriculum systems analyst to consider. The system proposed in this paper is capable of adjusting its operation anywhere along the continuum from batch to interactive computer usage. This feature enables one to study the relationship between cost and usage-mode in a systematic manner.

Another persistent problem concerns time demands made on the teacher by an instructional management system. It is interesting to note that in the preface to his arithmetic textbook, Cocker (1678) makes the claim that the book will serve as a monitor to instruct the young, enabling the teacher "to reserve your precious moments which might be exhausted that way, for your more important affairs." Ever since, there has been an incessant and largely unsuccessful struggle with the problem of saving the teacher's time. The shortage of instructional time for teachers remains critical today. In fact, rather than reducing the time demands, evidence seems to point to an increase in the time required of the teacher when a computer-based instructional system is introduced. The prototypic system suggested in this paper is designed to ease this critical problem rather than aggravate it.

The teacher time problem grows out of another system consideration, namely system stability. Research by Suppes (1968) and others has shown that individuals vary greatly in their natural learning rate. Allowed to proceed freely, learning groups that start at the same point in the curriculum and which begin as a homogeneous group rapidly diverge into smaller and smaller subgroups which in the limit become groups of unit size, i.e., each individual ultimately differs from every other individual. The management of such a system has so far showed itself to be beyond the capability of even the largest present day computer to handle.

The traditional classroom teaching system, however, is unstable in a different manner. Slow learners face continual failure experiences and eventually lose hope for success. The slow student often falls into a hopeless pathological state of mathematical ineptitude. Only the fittest pupils are able to survive the well-intentioned, but inadequate techniques of traditional group instruction. This system operates to maximize administrative utilities at the expense of individual learner utilities. It is intended to show here how one can adjust a management system to operate at various points along the "individual-group" continuum in keeping with practical constraints.

One of the most important and difficult problems for the curriculum analyst to deal with concerns the matter of change.

Curriculum models traditionally have been static models. As Koerner (1963) and many other critics have pointed out, the school system is ideally designed to resist change. Yet in areas such as the mathematics curriculum, academicians have sharply pointed out that change is desperately needed (CCSM, 1963). Thus it remains a problem for system analysis to find ways of developing dynamic curriculum models.

The work of Davis (1965) in the Madison Project illustrates one possible approach to the construction of dynamic curricula that he calls the hypodermic technique: new segments of the curriculum are polished and perfected and then "injected" into an existing curriculum. Another example is the IMS under development at SDC which is being designed as an evolutionary system projected ultimately to grow into a relatively complete program of instruction. These both operate largely on a command system orientation. The dynamic feature of the management system discussed in this dissertation lies in its potential for clearly diagnosing and isolating specific learning problems. It is equally adaptable to either the command or demand mode of instructional system operation. Basically, the mechanism for change is based on the assumption that if teachers have a clear picture of the learning problem an individual or group may have then it will be easier to make a convincing case for changing the existing instructional treatment in suitable ways.

1.6 Problems of Educational Measurement

In traditional instructional systems, curriculum planners set some general goals, students are provided with more or less the same instructional treatment, and then a test designed more or less according to classical principles, often by a third party, is used to determine the relative distribution of students as they are scattered across the field of attainment. Evaluation of short term learning is difficult at best in this situation. Bormuth (1968) asserts that "achievement tests, as they are currently constructed, cannot be claimed to have any objectively demonstrable relation to instruction." Wright (1967) less gently asserts that conventionally constructed tests "are no damn good."

One alternative under study is to design group instruction more carefully and devise methods of test writing which objectively measure criterion attainment for fixed instructional treatments (command mode operation). The work of Bormuth (1968), Hively (1968), and Coulson et al (1968) is directed toward this end.

Another option is to specify sets of problem situations in which, after instructional treatment, the pupils must demonstrate their capability of successfully solving a minimum percentage of the set. Where the population of items is large, the item pool is randomly sampled and the "true" proportion is estimated from performance on the sample.

This third procedure is the approach to be followed here. It is assumed that learning goals can be related by a curriculum system of stands, units, and topics specified at functional, foundational, and formal levels. Each topic within a unit specifies categories of problem situations that individuals may select for work toward a specified level of skill and concept mastery. It is the responsibility of the instructional designer to provide suitable individual and small group learning opportunities which maximize the probability of the pupil's attaining mastery in available time.

This procedure differs from other approaches in that one does not begin with statements of behavioral or operational objectives and then proceed to develop instructional strategies and write suitable evaluative items. Rather, content analysis leads first to the specification of significant problem categories from which one may, if it is desired, abstract a statement describing the apparent behaviors involved (DeVault and Kriewall, 1969, Chapters 3-5). Test design, from this point of view, is therefore more dependent on adequate content analysis by subject matter specialists than it is on item analysis by psychometrically skilled persons. To use the two-span bridge simile introduced by Cornfield and Tukey (1956), the approach suggested here represents a strengthening of the "subject matter span" of measurement possibly at the risk of weakening the "statistical span." The crucial concern is to provide options needed to adjust effectively between the often conflicting values inherent in subject matter and measurement considerations.

It perhaps needs to be stressed that no claim is made here that all learning can be evaluated against a mastery standard. This is why we distinguish, for instance, between the functional, foundational, and formal levels of mathematics learning. In general, pencil and paper test instruments can be designed more easily at the lower cognitive levels (Bloom, et al, 1956). That is not to say that one could not specify categories of theorems to prove, for example, to test both proficiency and elegance of proof at high cognitive levels. It is simply that practical constraints of money and time together with smaller probabilities of success limit the utility for doing this.

1.61 The Need for New Evaluation Metaphors

The trend toward individualization of instruction has forced changes in many educational values and practices. Traditional testing and grading practices, however, have not been readily adapted to certain instructional innovations recommended in recent years.

One of the evaluation problems faced by those concerned with individualization of instruction is that the classical norm-referenced test (NRT) is built, to use MacDonald's (1965) term, on a "mythology" that is inapplicable or irrelevant to many new instructional problems. In explanation of the term "mythology," MacDonald says:

> . . . we may utilize many metaphors in our talk about instruction. Some of these metaphors have been raised to the level of myths. They are myths by definition here because they are used to prescribe patterns for instruction--when in reality they are only possible ways of viewing, with uncertain probabilities of validity.

In much the same sense of the term only applied to measurement rather than instruction, we are suggesting that new metaphors are needed to clarify some evaluation problems which, as Glaser (1963) has indicated, have been clouded over by an entrenched NRT mythology.

Test metaphors usually arise as rationales or interpretations for procedures and assumptions initially adopted mainly on theoretical grounds. For example, the proportion of examinees who correctly answer an item is a well-defined theoretical construct interpreted as the "difficulty" of the item. A defining statement such as $\pi_g = \mathcal{E}_a(Y_{ga})$ has been called a *syntactic* definition (Lord and Novick, 1968, p.15). An empirical, behavioral, or *semantic* meaning such as "item difficulty" is what Carnap (1950) has called the *explication* of the construct. The term metaphor is used here because a change of context can render a given explication invalid. Metaphors, raised to a level at which they become an almost unchallenged basis for prescribing test construction procedures when in fact other alternatives may be just as or even more useful, are called myths. General and uncritical acceptance of myths leads to faulty test construction and confusion. The hypothesis defended in this argument is that the myth of classical test construction which prescribes item selection procedures based on consideration of constructs such as item difficulty, item validity, item discriminating power, item intercorrelation, and item-test correlation is not relevant to some important situations of interest to the instructional manager. Other means of test construction are not only possible but are very likely to produce more useful measures for management purposes. While all this is in some sense obvious, it is not difficult to find instances of research in education being guided by an accepted, conventional rhetoric rather than by appropriate inventions that may appear to challenge existing mythology.

1.611 Item Difficulty

This construct is defined as the expected relative score on an item by a population of examinees. It is often denoted by the symbol p (or pi) because of its interpretation as a probability. If an individual is selected at random from the population of examinees, then $\underline{p}$ is the probability such a person will respond correctly to the item. A difficulty with item difficulty, from the instructional manager or teacher's point of view, is that at the local level one is not teaching a random sample of children selected from a specified population but rather a particular group of individuals. Inferences must be made concerning these nonrandomly selected individuals. It is cold comfort to have a decision system that is right on the average if it were repeatedly applied to the hypothetical pupil population but wrong in every individual case at hand. Thus it is important that the individuals be treated as such and not as a random sample from some larger population, at least in the context of day-to-day instruction.

This means that, in this situation, item difficulty is not an appropriate or particularly useful concept in its classical sense. A new metaphor is required. Now the goals of instruction can frequently be cast in terms of developing problem solving behaviors. At a given point in time, the teacher is trying to develop specific sets of problem solving behaviors. The pupil may develop the desired behavior in various ways. He may completely fail to comprehend the ideas involved. Or he may develop an algorithm which works on some problems but not all of a given kind. Or he may learn adequate, general procedures that render all problems of a given class equally capable of solution subject to the normal human failures brought about by random personal or environmental sources of error. What the teacher needs to know at given points in time is what this probability for success is for a given pupil with respect to a specified class of problems. Rather than sample problem solving behavior across a hypothetical population of pupils, it is more appropriate to measure the individual's behavior on a random sample of problems drawn from a clearly defined population of problems. The individual's relative score on this sample can then be interpreted as an estimate of his proficiency relative to that class of problems. This metaphor not only clarifies what we might intuitively regard as the difficulty of a given kind of problem for a particular child but also helps define the objective of instruction in terms of the level of proficiency expected in the learning product. The higher proficiency is, the better the quality of the educational system's output.

This approach to test construction has been receiving increased attention in recent years (Hammock, 1960; Ebel, 1962; Glaser, 1963). We shall call tests constructed to provide proficiency measures, as described above, criterion-referenced tests or CRTs. A model for such tests will be developed and compared with existing models of interest in Chapter II. Uses for the measurement data produced by CRTs will be indicated in connection with the instructional management system described in Chapter III.

1.612 The Assumption of Normality

Another metaphor that is commonplace in classical test construction is that tests measure one or more mental traits and that these traits are distributed among a population of examinees according to the normal curve of error. Given mass data, the supposition appears to be valid for many mental traits of interest. However, the assumption of a normal distribution for proficiency in particular problem solving skills in a given classroom is, at the least, suspect. It seems a good deal more likely that proficiency distributions are bimodal rather than normal. For some pupils, the needed skills are pretty well understood in sufficient generality to permit frequent success on repeated random trials. For others, procedures may have been learned which work for certain special cases but not in general. Such students would be expected to show a gradually improving but excessively limited level of proficiency. The data of interest to the teacher are not the class mean and relative ranking of class members but rather the correct classification of students into mastery and nonmastery groups together with estimates of absolute levels of proficiency within each level. These data would be sufficient to assist the teacher in making important instructional decisions in regard to differentiating instruction and in comparing the effectiveness of alternative instructional treatments.

The list of malpractices developed in connection with the normal distribution myth is a testimony to the hazards involved in the uncritical acceptance of metaphors. Traditional single-treatment group instruction does produce score distributions that are often close to the normal curve. The reason is that a single treatment for a specified period of time is not equally appropriate for individual learners. Some need more time, others less; some need more explanation, others less; some prefer to discover their own conclusions, others learn better by preceptive or didactic strategies. The distribution of scores according to the normal curve of error reflects "error" in instructional design and treatment. The malpractice comes about in grading on the curve for it assumes that the source of "error" lies in differing natural gifts and abilities rather than inappropriate instruction. The test results are then misused so frequently by grading the quality of these presumed natural abilities on a scale from A to E, much like one would grade the quality of eggs.

A second malpractice motivated by the normal curve myth is averaging grades over time. Ordinarily, one repeats measurements and averages them in order to cancel out the errors of measurement which are assumed to be normally distributed about the mean. Teachers often accumulate a series of test grades over a period of time then average these to obtain an estimate of the pupil's true ability, presumably on the assumption that the average is closer to the true value than any single measurement because the errors are canceled out this way. Such nonsense is deeply imbedded in existing grading practices largely because of the uncritical acceptance of the normal distribution myth and its implications.

Individualized instruction attempts to compensate for natural learning differences by differentiating the form of instruction to suit individual learners or small groups of learners having similar instructional needs. Time given to instruction also is adjusted as the need requires.

Once the class is divided into two or more learning groups, the traditional norm-oriented grading system breaks down. How does one compare performance of individuals in different, non-comparable learning groups? Should A's be given only to the best individuals in the "high" group? Or does that predestine slower learners to D's and E's? If a slow learner takes four tests on a given learning objective and fails all of them, then "catches on" and gets an A on the last test, how should his achievement be recorded: as an A, because now he really knows what he is about; or as the average of four E's and an A on the assumption that the A might be due to an error in measurement?

Chapter II describes means of measurement that are suited to the requirements of assessment in individualized instructional settings. Assumptions of normal distribution are neither needed nor used. Hopefully, the measures generated on the basis of the model suggested will not only improve techniques of instructional management but help eliminate practices such as those cited above together with others of equally dubious merit that are built on the assumption of normality.

1.613 Test Reliability

Reliability is defined in classical test theory as the squared correlation between true and observed scores. The metaphor used to give semantic meaning to this syntactic definition is that reliability measures the extent to which a repeated measure would agree with the original measure on a group of examinees. If the examinees' traits measured by the test have not changed, then another administration of a test which measures the same traits should ideally provide the same score for each examinee.

The metaphor becomes a myth when it is used to prescribe methods of test construction. It is easy to show that maximum variance is achieved when item difficulties are approximately .50. Thus, for maximum predictable test reliability, it is commonly recommended that use of items with either very low or very high p-values be avoided (Lord and Novick, 1968, p.329). The problem with this procedure, from the teacher's point of view, has been already indicated (Sec. 1.611). The "difficulty" of items for a nonrandomly selected group is, first of all, not known before the test is administered. In fact, it might be considered that the purpose of the test is to gain information about the item difficulties as a measure of instructional effectiveness inasmuch as the test scores and item difficulties are functionally related (expected test score for a group of examinees equals the sum of expected item difficulties). Furthermore, instructional objectives could well be defined by identifying the classes of items for which the "difficulty" will intentionally be reduced as a result of successful instruction. From the instructor's point of view, this presents a constraint which conflicts with classical prescriptions for reliable test construction.

Suppose the teacher is trying to develop a specific set of problem-solving skills such as transforming sentences with active voice to ones with passive voice; or, in arithmetic, with teaching methods of solving relative motion problems. The pool of questions which can legitimately be asked on a post test are determined by these instructional objectives and not by measurement requirements involving the consideration of item difficulties, variance, or intercorrelation. It is conceivable that the instruction may be either effective or ineffective. In either case, score variance will be relatively small and hence the reliability estimates such as KR-20 would likely be near zero. Certainly there would be no point to restructuring the test simply to produce more variance. Such a procedure would be like trying to improve the quality of production by setting more stringent acceptance requirements in place of improving the "production", or instructional, techniques.

Part of the problem arises from the classical aversion to the use of rawscore as a meaningful measure (Gardner, 1962). The teacher, however, is not so much interested in the relative ranking of pupils as in knowledge about their absolute levels of proficiency in problem-solving situations. Tests designed to make the absolute score meaningful are therefore required to meet this need. Prescriptions for constructing such tests involve both a different model and different metaphors from those of the classical model. These differences are treated in the following chapter in detail.

CHAPTER II

MANAGEMENT SUPPORT SYSTEMS: THE MEASUREMENT INFORMATION COMPONENT

2.0 Introduction

The operation of an instructional management system is dependent for its success upon the effective functioning of certain support components. One of these subsystems might be called the measurement information component. The function of a measurement component is to generate learning data that can be used for any one of several purposes. These purposes may include use of measurement data to (1) categorize learners into groups on the basis of their common requirement for instructional treatment (Diagnosis and Prescription Function); (2) to assess the relative effectiveness of competing instructional treatments (Instructional Assessment Function); (3) to determine, in the case of established instructional segments having predetermined performance standards, which individuals have acquired minimal standards of proficiency required for mastery and which learners require further prescriptive assistance (Quality Control Function); and (4) in the case of curriculum development, to indicate hierarchical relations within a content sequence (Curriculum Design Function).

This chapter discusses a model for a measurement information subsystem which generates data by using criterion-referenced tests of a particular genre.

Chapter III shows how the model can be used for applications such as (1) and (3) cited above, while Chapter IV summarizes implications of the model for further research in other related areas, such as (2) and (4).

2.1 Characteristics of Criterion-tests

2.11 Absolute vs. Relative Measures

A teacher's use of test data falls roughly into three categories. He wants to know, at frequent intervals, what the members of the class have learned during the intervening period of time; he needs to gather data for grading purposes; and he is often interested in the progress of the class relative to national norm groups. If the teacher is skilled at test construction, he is likely to build tests for grading purposes along classical lines, that is, in such a way as to maximize differences among individuals. The standardized tests secured for end of semester assessment and norm-group comparison will also be designed to maximize individual differences. The important consideration in the design and use of these tests is the reliable ranking of pupils. Therefore the measures generated by the test reflect relative standing in some general content or skill areas rather than absolute levels of achievement within some specific body of content. While useful for summative evaluation, these kinds of data have not been found to be very useful for short-term instructional decision making (Coulson *et al* 1965, 1968).

Rather, what is required to guide instructional decisions is curriculum-specific data that can be meaningfully interpreted in the absence of pupil-group data. Hence one needs absolute measures of an individual's proficiency with respect to a well-defined body of content or set of skills.

Tests which provide relative measures are commonly called norm-referenced tests; tests which provide absolute measures are termed criterion-referenced, or simply criterion tests. As is the case with norm-referenced tests, many models have been devised to guide test builders in the construction of criterion tests. Most of these, as Glaser (1963) has indicated, often reflect the influence of the more firmly established item selection procedures characteristic of norm-referenced tests. It is the intent of the argument of this chapter to establish a useful and flexible model for criterion-test construction, one which provides for the generation of content specific data to be used in making instructional decisions on individual cases. The absolute measure so generated is referred to as "proficiency" in the following discussion.

2.12 Content Validity

What mental traits a test measures is a question which the psychometrician has no adequate way of answering (Lord and Novick, 1968, p.528). It is because of this haziness that classical item selection procedures serve only as a guide rather than an algorithm for test construction. In the final analysis, the test builder must make subjective decisions concerning a given item's relation to whatever it is he wants to measure.

However, the usefulness of a criterion-test is vitiated unless the test has obvious content validity (Ebel, 1962). It is of little use to an instructional manager to know a pupil is 90% proficient, for example, if it is not known what specific content or skills compose the proficiency. The model to be developed here, therefore, will be developed with the need for prima facie content validity in mind. The essential construct that enables one to meet this condition is the notion of a well-defined item population or "specified content objective", as it will be denoted in what follows. This represents a different approach from classical test construction in two important ways.

The first, the most crucial, departure from classical test construction will be seen in the fact that the usual item selection procedures are not, and indeed cannot, be employed. The second point of difference lies in the manner that one relates instructional objectives to evaluation strategy.

It has become conventional, in principle at least, to insist on behaviorally or operationally formulated statements of instructional objectives. The usual strategy for evaluation then involves the translation of such statements into test items which determine whether in fact the desired capacities for new behavior have been developed. The strategy involved in the design of criterion-tests on the basis of the model proposed here begins, not with the specification of expected student behavior, but rather with the specification of a problem solving category with respect to which proficiency is to be developed. This category of problems is called a specific content objective.

This approach accomplishes two important things. First it provides an algorithm for test construction and thus secures the content validity required. Secondly it provides, through the test, a proper beacon for instruction.

It is well known that teachers have a propensity for teaching "to the test." Usually this practice is discouraged and for good reason. If a test contains a collection of items on a variety of topics, teaching to the test destroys the usefulness of the results for the purpose of norm-group comparison. One can conclude from the test results very little except the degree to which the students are capable of retaining crammed-in bits of unrelated knowledge.

However, if the test is constructed by random sampling from a specific class of problems (e.g. situations requiring grammatical analysis, computation, or disciplined patterns of reasoning), then it is desirable that the instruction be directed to the development of relevant skills in sufficient generality that a uniformly high probability exists for successful behavior no matter which particular problem is selected for test purposes. It is in this sense that the criterion-tests discussed below are intended to be appropriate beacons for instruction.

2.13 Parallel Item Selection

The dependence of classical test construction on subjective decisions made by the test builder (Lord and Novick, 1968, p.350) is wholly undesirable in criterion-test construction not only because of its deleterious effects on content validity but also because of a need that exists in individualized instructional systems for the generation of many "parallel" tests (Hively _et al_, 1968). The instructional paradigm depicted in Figure 2, Appendix G implies the possibility of an individual repeatedly recycling through a given body of content (usually over an extended period of time as other learning goals are interleaved). Upon the completion of each cycle, no matter how the sequencing problem is handled, one needs a new version of the test to determine if the pupil has finally achieved some minimal level of mastery.

An obvious technique to investigate for the purpose of generating many parallel versions of a given criterion-test is the use of the random number generator of a computer to sample the specified item population. In particular content areas, such as mathematics, it is further intriguing to consider the possibility of actually generating the items in random fashion rather than randomly recalling them from a prepared list. The criterion-test model proposed in this paper is purposely designed to facilitate the use of the computer item-generation technique. However, it should be noted that this is more a matter of convenience or facility rather than a restriction in the range of usefulness of the model. The agent used to generate the tests is only incidental; the method used, however, is critical. In the case of computer generation of items, the only question is whether or not in the particular case at hand the computer is a more efficient or economical agent than any other that is available.

2.14 Minimal Test Length

Consideration of the paradigm in Figure 2 suggests the possibility of a conflict arising between the value systems of the instructor and the evaluator of instructional effectiveness (even if both functions are performed by one and the same person). This instructional model, which is common to many current individualized instructional systems, involves pre-test, instruction, and post-test. If, as is the usual case, a fixed amount of time is available for the combined functions of instruction and evaluation, then the allocation of more time to one function necessarily decreases the time alloted for the other and a conflict exists. The instructor presses for more time in the hope of achieving higher levels of learning while the evaluator requires more time to either sample a greater range of specified objectives or to get better estimates of proficiency on a given selection of objectives (e.g., Walbesser and Carter, 1969).

The problem for the systems analyst is to find an admissable, if not an optimum, solution to the conflict. The criterion-test model to be described is designed to be useful in the solution of this problem in two important ways.

The two viable options open are, first, to reduce test length while preserving efficiency* and, secondly, to use convergent testing strategies.+ The former can be handled through the test model in an analytical fashion; the latter by competent content analysis. Although the content analysis is in large part a judgmental matter, the test model helps focus attention on the central concerns by virtue of its emphasis on specified content objectives. Consideration of the problem of minimizing test length leads to the use of acceptance sampling theory and methods of curtailing inspection, such as Wald's Sequential Probability Ratio Test (SPRT).

2.15 Criterion Selection

The term "criterion" is often used in measurement terminology to denote a predicted variable, particularly in discussions relating to the question of classical test validity. In this paper, however, a criterion means either a cutting score or a limiting value of a proficiency range. For example, on a five item test one might set an error criterion of 2 so that pupils who have 0 or 1 errors are classified into one instructional group while those having 2 or more errors are classified into another instructional group. A similar but formally different illustration involves hypothesis testing. Suppose one wishes to classify learners into a high or low proficiency category. The extreme limits of proficiency are determined naturally: those who always get every problem right are obviously masters and those who always get every problem wrong are nonmasters. But it is also reasonable to allow for some variation in behavior so that the mastery range might extend from perfect performance, p_0, (zero error rate) up to some value p_M, for example, which denotes the maximum proportion of errors allowable in the range of performance definitely considered as mastery. The value p_M serves as an upper bound to this proficiency range. Similarly, nonmastery may be defined to include the range of proficiency from 100% error rate, p_1, down to some value p_N, the least error rate definitely considered as an indication of nonmastery. The values p_M and p_N are criterion values used in hypothesis testing associated with the "Quality-control Function" mentioned earlier.

---

* "efficiency" is taken to mean "having adequately small probabilities for all relevant kinds of errors." (Birnbaum in Lord and Novick, 1968, p.436). +A convergent strategy depends on the existence of an inclusion relation between the ability being assessed and component abilities also of interest. For example, long division requires subtraction and multiplication operations to be performed in succession. Therefore, the measure of success on a test of division proficiency is a lower bound to the separate component proficiencies of multiplication and subtraction. Thus if long division proficiency is high, one can infer that both multiplication and subtraction proficiency are at least as high. The converse is not true, however.

This raises the question of how one selects criterion values. A survey of existing systems indicates a tendency to specify a rigid criterion selection policy. Usually criteria are indicated by stating percent values such as 80%, 90%, or 100% as the minimum acceptable level of mastery performance. Analysis of sampling plans is rarely performed and one often finds little attention given to the decision-implications inherent in the casual selection of test length together with a fixed-policy criterion. It is not difficult to find instances where higher criteria are selected in the mistaken belief that this will result in a better quality of learning product than will a system having a lower criterion.

The model to be developed will yield methods for analyzing the complications of given test length and criterion value selections. It will be shown that, in general, the proclivity to fixed criterion usage is not likely to be an adequate policy. Rather, the discussion of sampling plans will indicate that both the criterion and test length need to be selected in a way suited to the context in which the test is to be used.

In summary, the measurement information component envisioned here is designed for use in instructional management systems where classifications of pupils for treatment are to be decided on the basis of minimal data consistent with predetermined limits for the errors of misclassification. The measures obtained are content-specific estimates of proficiency useful for the stratification of learning groups on a day-to-day basis if need be. By sampling across items rather than across persons, absolute measures of proficiency are obtained which can be reliably interpreted for non-randomly selected pupils, the pupils of particular instructional concern. The model is designed for wide variety of applications but retains in the concept of proficiency a simple semantic explication. The empirical data generated are intended to have clear although not necessarily causal implications for instructional decision-making.

2.2 A Heuristic for Criterion-Referenced Tests

Criterion tests having the properties discussed in section 2.1 can be constructed using a strict item sampling model. The term "strict" simply means that one first defines the item population, then selects a random sample of $\underline{n}$ items for test.

This rather obvious procedure is emphasized because it is at variance with conventional item-sampling techniques. Cornfield and Tukey (1956) have characterized the more usual approach as one involving first the choice of a sample on which statistical analyses are made then introducing an unspecified population of items "like those observed" for which inferences are to be made. With the same perspective, Lord and Novick (1968, p.234) speak in terms of "$\underline{n}$ test items considered as a random sample from a population of items", rather than $\underline{n}$ items which are a random sample. Loevinger (1965) and others have criticezed the item-sampling model because tests are not constructed by actually drawing items randomly from a specified population. The model proposed here avoids this criticism by specifying a method of selecting items which indeed compose a random sample from a well-defined item population.

It should be noted that this strict sampling procedure is introduced for other reasons than simply to blunt criticism such as Loevinger's. Alternative rationalizations for the "opposite" approach to item sampling have been constructed in much the same vein that one defends the concept of parallel tests in classical test theory. Thus it is not the intention here to enter the controversy between psychometricians with regard to the item sampling model, but simply to note that objections have been raised against it and that, by incidentally meeting them, a distinctive test model is being used.

2.21 Assumptions of The Item-Sampling Model

Assumption 2-1: There exists, for each criterion-test, a population of tasks which can be specified in set-theoretic terms.

What the psychometrician calls a population, the mathematician might call a universe or a universal set. Sets, in mathematics, are well-defined collections of objects. "Well-defined", in this usage, means that decision rules can be composed which unambiguously describe the attributes that elements in the set must possess. Therefore, given an item, it is possible to decide whether it either is or is not a member of the set.

Assumption 2-1 asserts that a relevant situation for the use of a criterion-test is one in which it is possible to define, *a priori*, a universal set or population of items. For example, it may be desired to test a pupil's proficiency in detecting whether or not a pair of randomly selected three-letter (nonsense) words are the same, where each pair of words is built on the pattern "consonant-vowel-consonant." One might further restrict the first and last consonant to have certain properties such as being the same within a word but randomly different or the same between pairs of words being tested. The "replacement set" from which the consonants and vowels are to be randomly selected can be specified as desired. In this way the set of tasks to be tested becomes well-defined and, by random sampling from the population, one can (1) estimate an individual's proficiency relative to the defined task population or (2) on the basis of minimum test size and specified limits of classification errors, classify individuals into groups which (a) have proficiency greater or equal to some minimal mastery criterion or (b) have proficiency less than or equal to some maximum nonmastery criterion. In general, it is not possible to achieve efficiently both functions of accurate estimation and economical classification with a single test administration. This conflict will be discussed later.

A label adopted here to denote the well-defined item population described above is the "specified content objective" or SCO. The item population is said to be specified, rather than specific since the use of the term "specific", as in "specific behavioral objective," leads to confusion. The confusion results from ambiguity concerning how specific an objective must be to be called specific. An objective is specified when the rule or procedure which defines membership in the population is clearly stated.

Assumption 2-2: Each pupil has a single proficiency at any given point in time relative to a specified content objective.

Assumption 2-3: Proficiency is an increasing function of instructional time.

These assumptions, like the first one, serve to place restrictions on item selection procedures. Assumption 2-2 in particular implies that one cannot successfully employ the global item selection techniques associated with classical test construction. The classicist is content to talk in rather vague, general terms of such constructs as arithmetic ability, geometric ability, special ability. The instructional manager and teacher need to have diagnostic measure on a much finer level of detail. This need for refinement suggests that the population of items be restricted by specifying content parameters, such as have been suggested by DeVault and Kriewall (1969, p.116).

Assumption 2.3 finds application mainly in Chapter IV where the predictions or implications of CRT Theory concerning the parametric specification of a proficiency learning curve are discussed. Actually, proficiency development is more likely to be a series of exponential growth and decay curves, dependent upon the degree of utilization which the pupil finds for particular concepts and skills over a given period of time. It is also possible that a given instructional segment may cause temporary degrading of proficiency levels, especially if a new method is introduced which interferes with some previously learned skill or concept. In any case, Assumption 2.3 says that these kinds of complexities will be ignored for the sake of simplicity in the following consideration of CRT Theory.

2.22 Implications for Homogeneous Grouping

It should also be noted that the concept of proficiency, *vis a vis* ability, is used to connote a dynamic, as opposed to static, learning parameter. Proficiency, when defined relative to an SCO, can be a rapidly fluctuating quantity. If the instructional treatment is successful, perceptible gains can be sensed and quantified on a short term basis. Furthermore, it appears possible that one can, with some accuracy, separate the components of proficiency and particularize the prescription for meeting observed learning problems.

Therefore, proficiency serves as a useful parameter for instructional decision-making. One application is to use it for forming homogeneous learning groups. The formation of homogeneous groups, it may be noted, is always possible both in principle and in practice. The trick is to specify the attribute(s) with respect to which the group is to be homogeneous. The group so formed should be described as being homogeneous with respect to the specified attributes. We often speak elliptically and simply refer to homogeneous groups.

The point to be made here is the following. As indicated below, schools have traditionally worked with homogeneous groups. The question is, not whether this can or should be done, but what is an effective parameter to use in forming homogeneous groups for instructional treatment? The answer proposed here is--proficiency. Since the late 1800's, we have stratified pupils on the basis of chronological age. Thus our school "grades" are homogeneous with respect to age. We know that such methods of stratification do not yield effective instructional groups. Therefore we look to other parameters on which to individualize instruction. An extreme parameter is the discrete individuality of a pupil; we can form instructional groups of unit size. For a variety of reasons this is impractical. We can stratify on IQ or achievement score values. Some evidence exists that this yields no significant difference in learning efficiency over chronological grouping. Furthermore, these are all relatively static parameters which do not sensitively reflect the short term changes in learning.

Proficiency, on the other hand, has the desired properties for dynamic, short term creation of temporary homogeneous groups for instructional treatment. The following illustration indicates, in rough outline, the procedure involved. In recent tests conducted by the author (see Appendix A), a group of 19 children were given a criterion post-test designed on the basis of the CRT item-sampling model. Three distinct proficiency groups were identified. Eight children were members of a "zero" proficiency group; six belonged to an intermediary proficiency group of about 45%; the remaining five were slightly above 90% proficient.

It is evident that these groups have distinct instructional needs. Although no causal data exists to relate a given level of proficiency with a particular instructional treatment, one might speculate that the top group would need no further instruction on the SCO at this time; the middle group might possibly need some minimal instruction and certainly more practice; finally, the low proficiency group would appear to require complete reinstruction. In the last case, it would appear that further drill could not be expected to serve the needs of the lowest proficiency group.

Following such classification decisions and subsequent treatment, one can observe the effects of instructional treatment in two ways. Some children make transitions from a low proficiency state (nonmastery) to a high proficiency state (mastery). The proportion of children who make this transition is one index of instructional effectiveness. The other datum reflects a change in the mean proficiency within the remaining homogeneous (with respect to proficiency) groups of learners (cf. Appendix A, pp 186-187. This approach gives needed semantic meaning to such syntactic definitions used in connection with educational product development as 90/90 (90 per cent of the class makes the transition to a mastery state defined by 90 per cent minimal proficiency). Thus in the following discussion, "homogeneous" will always be taken to mean homogeneous with respect to proficiency. Furthermore, proficiency is always understood to be relative to a specified content objective. These assumptions and constructs are formalized in the following model.

2.3 A Formal Theory of Criterion Tests

2.31 Functions of a Theory

A theory has three essential functions to perform. It must *explain* known phenomena, *predict* new phenomena, and *suggest* new ideas for research and development. In the case of the CRT Theory described below, "known" phenomena encountered in the development of instructional management systems have stimulated the ideas on which the theory is founded. Thus it should be expected to meet the first criterion to some degree.

With regard to the prediction aspect, a theory accurately predicts outcomes only for real world situations which are isomorphic to the structure specified by the theory. In this sense, the prediction function reduces to a truism: something is always predicted by a model. But we recall that a theory is only an approximation to real world situations and the risks one takes in the search for simplification only become serious if too much of the real world structure is ignored. The trick is to achieve sufficient simplification to make the model analytically manageable while at the same time retaining predictive validity in situations of practical interest. Therefore an interesting theory is not only logically tight but practically applicable. Although the theoretical development reported below is only a first order approximation to real world performance, it appears to be an interesting theory in the sense described above.

2.32 Order of a Model

By "first order" model is meant that the simplest seemingly feasible assumptions are made. For example, the assumption that all items have essentially the same "difficulty" leads to a single-parameter binomial model for criterion-tests. This assumption could be relaxed to permit consideration of the case where each item may have a distinct difficulty-value. This would lead to a n-th order model based on the compound binomial.

All discrete levels of intermediate-order models can be constructed based on the "mixed binomial" model having two or more parameters. If, for example, a specified content objective were composed of two distinct item-difficulty classes, then repeated testing by random sampling of n items from the SCO would produce score distributions given by the double-binomial, a second order model. In principle, one could fit any distribution with a mixed binomial of high enough order but as Hill (1960) notes, " . . . the whole thing becomes rather meaningless if too many components are taken." For example, proficiency as estimated by the mean of a first order distribution retains a simple pragmatic and semantic meaning. But what is to be understood by a proficiency measure which is the mean of two or more separate distributions? One could attempt to fit an observed distribution with a higher order model, but the problem of estimating the several parameters leads to trouble. As Lord and Novick (1968) report in regard to the compound binomial model, sampling errors destroy the utility of higher order models unless simplifying assumptions are made. The classical model might be considered in this view as an infinite-order model. The simplifying assumption needed to make the model workable is that of a normal distribution. This is a two-parameter model determined by the mean and variance. However, the cost of this assumption is virtually the total loss of content meaningfulness of the resulting measure. One simply obtains ranking on a semantically undefined scale. (Angoff, 1962)

Another dimension on which the model under consideration is restricted has to do with hypothesis testing, implications of which are developed more fully in Chapter III. When classifying pupils for instructional purposes, one may assume that only two distinct proficiency groups exist, or three, or more as desired. Techniques for solving the "two-way" decision problem have been thoroughly worked out (Wald, 1947, 1950; Wald and Wolfowitz, 1948; Statistics Research Group, 1945). When one goes beyond this point to "multiple-decision" problems, the literature reveals that many obstacles stand in the way of practical implementation (Amster, 1963; Anscombe, 1953; Armitage, 1950, 1960; Beckhofer, 1954, 1958).

As a start, therefore, it is well to begin with the simplest viable model and take into account the limitations when making inferences about the real world. One can subsequently step-up the order of the theory as need manifests itself and technique improves to handle the increased complexity.

The viability of a theory shows most clearly in the extent to which it suggests new ideas for test and new procedures for making sharp experiments. This aspect of the CRT theory under study is discussed in Chapter IV.

2.33 Basic Definitions for A first Order Theory

There are four fundamental definitions with which we begin;

in particular, the definitions for (1) a specified content objective, (2) proficiency, (3) error rate, and (4) criterion-referenced test. Definitions always involve semantic, syntactic, and pragmatic meanings. Therefore we will attempt to indicate not only the referent in each definition (semantic meaning), but also a context for appropriate usage (syntactic meaning) and brief descriptions of practical connotations (pragmatic meaning) wherever necessary to make the definition clear.

2.331 Specified Content Objective

The concept of a specified content objective develops directly from previously cited CRT applications, in particular, the assessment of absolute levels of proficiency on specified categories of problems. In effect, we are saying that it makes little sense to speak of proficiency in the abstract. For practical purposes of instructional management, we need to relate proficiency to a well-defined domain of tasks which define the content of the test. A sufficient method for insuring clear specification of problem domains involves the use of item generation rules. We therefore define a specified content objective as follows:

Definition 2-1: A specified content objective (SCO) is a rule or procedure for generating a class of problems.

The word "problem" is used in the definition to connote the idea of problem solving situation. Most frequently in applications, the "problems" will take the form of written test items, but this is not a necessary restriction. The essential feature of the definition is that a content objective is "specified" only if one can state a rule or procedure for generating the entire population of items that is "tied to" the content objective. The definition does not require that all items actually be generated or listed, nor is it necessary to compute how many items comprise the population in most instances.

There are, or course, many questions concerning the nature of such rules which we could raise at this point. We note only that it is a current topic of research interest. Bormuth (1968) approaches the use of generating rules via transformational and structural grammar. Hively (1968) reports the use of generating rules in connection with the mathematics testing of Job Corps trainees. Although requiring further research, suffice it to say here that the concept of using item generation rules appear to be practically feasible as well as theoretically useful.

Examples of a variety of item-generation rules can be found in Appendices, A, B, and F. In Appendix A, for example, the first test consists of items involving reduction of common fractions to lowest terms. The whole numbers for the numerator and denominator are generated by random selection of prime factors from the set {2,3,5,7}. Each pair of numbers comprising a given fraction was required to have in common at least one factor and both numbers were required to be positive and less than 100.

A lengthy description of the item-generation rule, such as that given above, can be considerably shortened by parametric specification of the rule. The data file named STAN/CMS, found in Appendix F, shows how item-generation rules are specified for tests 1-19, included in Appendix B. The procedure used in these cases was to work from a verbal description, such as that given above, to an identification of parameters which completely specify the item population. The generating program, called MATH/CMS (cf. Appendix F) then reads the item-specification parameters and generates suitably modified random numbers which conform to the rule.

2.332 Proficiency and True Score

In order to arrive at a convenient definition of "proficiency", we imagine the curriculum to be structured in terms of some network of SCO's. For example, the Strand-unit organization mentioned earlier in connection with DeVault's IMCP could easily be recast in such a form. Hively's structure resembles a PERT diagram; if Gagné's notion of hierarchies is valid, the structure may be made to resemble a decision tree (cf. Hunt et al, 1966). Whatever the most appropriate form for a curriculum structure may turn out to be, we suppose that one can be built and consider generic element of the structure called SCO_k, where k serves as the labeling index. Now suppose that student #a has completed some phase of instruction with respect to SCO_k. Further imagine that we require the student to respond to all the items in the population of items tied to SCO_k by the generating rule. The expected proportion of items to which the student exhibits a correct response is the measure of his proficiency which we shall use in the following discussion.

Definition 2-2:

The proficiency of the a^{th} student with respect to the k^{th} SCO,

> denoted by the symbol ζ_{ak} is defined to be the relative true score of *a* on all n_k items, (i.e., proportion of correct responses that individual #*a* would display if he were to respond to the entire population of items in SCO #k).

We observe that in the statistical sense of the term, ζ is a *parameter* or "population value." It is also, by virtue of being a parameter, a *constant* value for a given item-population and individual, at a given point in time. Pragmatically, therefore, proficiency may be taken to mean *the fixed probability of a correct response to an item randomly selected from the k^{th} SCO's population of problems*. In the same sense that the Π-value serves as a measure of a given item's difficulty for a pupil-population, ζ_a may be regarded as the mean difficulty of an item population for a given individual.

The complement of proficiency is termed the "error rate."

Definition 2-3:

$\zeta'_{ak} = 1 - \zeta_{ak}$ (error rate)

In our discussion of sampling plans later in this paper, it will turn out that the test length and criterion value are functions of proficiency. There is a strong analogy utilized at that time between industrial quality control analysis and CRT sampling theory. Industrial analysts show a preference to talk in terms of "rejects", or proportion defective, found in a batch being sampled (much as we might speak of an individual pupil's errors in a population of item responses being sampled) rather than in terms of "successes" found in the batch. The literature therefore usually shows equations and graphs expressed in terms of the error rate, to use our terminology, or the "proportion defective" in industrial terms. Although we prefer not to use the industrial terminology for obvious reasons, it nevertheless seemed appropriate, in spite of the negative connotation, to adopt "error rate" as the independent variable in order to make comparisons with quality control theory easier to follow.

It is further important to note that this is at once an important strength and weakness of the proposed CRT theory. Although there has been much talk in recent years about applying the seemingly successful management techniques of business to educational problems, it has not been easy to find useful educational analogies to the objective entities which business management worries about, such as articles of production (output); dollar cost; storage, shipping and distribution cost; overhead, and the like. It is a modest but important step that a definition of proficiency can be formulated that bears sufficient relationship to industrial production concepts to enable the utilization of quality control techniques described in the next chapter. The weakness lies in the possibility that we may be stretching a point beyond reasonable application. Only time and some carefully planned experiments will tell for certain.

The definition of proficiency as a true score raises a question of what is meant by the term "true score." At least three kinds of true score definitions can be found in the literature. There is, for example, the von Mises or "relative frequency" concept which defines true score as the limiting value of the average of $\underline{n}$ repeated measures on an individual as n goes to infinity (Carnap, 1962). Sutcliffe (1965) has introduced the notion of a Platonic true score, a construct that assumes that there exists some unique, natural true value for every individual on the measure being taken. Finally, there is the axiomatic true score defined as the expected value of observed score subject to certain conditions relating true score, observed score, and error of measurement, as in classical test theory. Axiomatic true score may appear in one of two forms, specific or generic (Lord and Novick, 1968).

The definition of proficiency given in definition 2-2 corresponds to the generic axiomatic construct. It is the expected score over a population of nominally parallel test forms which are obtained by random selection from a well-defined population of items.

2.333 Criterion-referenced Test

The true score, as a population value, is an unobservable quantity. One needs a device to measure or estimate this value. The measuring instrument is a criterion-test. From the definition 2-3 of proficiency, we draw the following definition of a CRT:

Definition 2-4: A criterion-referenced test is a random sample of items selected without replacement from an SCO.

Information theory tells us that a measurement, any measurement, can be reduced to counting a sequence of binary decisions on successively finer scales of measure. Error occurs when, for any one of a variety of causes, the count is improperly executed or prematurely terminated. The sources of error can be traced to one or more of the following causes. For example, where scales are involved, as on a meterstick, counting is effected by the use of marks on the scales. But the marks may be improperly positioned during manufacture. Errors of measurement arising from such a source are known as calibration errors. In other cases, a poorly designed measurement procedure may cause a fixed amount to be added to each measure. Errors which cause all measurements to be off by the same fixed amount result from what is called bias in the measurement. These are both sources of errors which can be minimized and possibly eliminated through careful methodology. There are other sources of error which can be reduced but never eliminated from measurements due to (1) interaction of the measurer with the thing being measured and due to (2) "noise" or random factors. These unavoidable sources will be treated later.

First we consider calibration errors and bias. The criterion test is an instrument for measuring proficiency in much the same sense that a meterstick is an instrument for measuring distance. The analogy is worth pursuing in some detail. Measurement of a distance, like any other kind of measurement, is a process consisting of a series of binary decisions. This fundamental method underlying all measurement can easily be understood by considering the process of measuring, for example, the length of a table. One brings up a unit of measure and asks "Is the table longer than this unit?" If the answer is yes, a count of 1 is recorded and a second unit of like size is brought up or the first unit is repositioned in the place of a second unit. The question is repeated: "Is the table longer than 2 units?" If the answer is yes, a count of 2 is recorded. The process then continues, each time the count being incremented by 1 until a NO decision is reached. Next, one repeats the sequence of questions using a subunit of measure, say 1/10th of a meter. The count of YES decisions is recorded in the tenths place; and so on until finally one reaches a sufficiently small unit of measure that the irregularities and rounding encountered at the edge of the table make further decisions impossible. We note for subsequent reference that the counts at different levels of refinement are first coded by a power of ten before being added together. (One does not combine 3 meters with 2 decimeters to report a count of "5" as the measure).

This illustration points up a number of useful generalities concerning the nature of measurement. The natural unit which is common to all measurement is the binary decision. The information contained in a measure, such as 2.732 meters, is the number of binary decisions made. The record of the measurement, to be meaningful, must contain both a numeric and a unit. To say a table is 6.51 long, for instance, conveys no information of value at all. Finally, the measurement is only approximate because at some level of refinement it becomes impossible to make a reliable decision. Thus, every measure contains errors to some degree from one or more sources.

2.34 Sources of Measurement Error

There are four sources of error of particular interest in obtaining criterion-measures. In the example above, if the zero point is not properly lined up with the initial edge, a fixed amount of error is added to the measurement. Such error is known as "bias". Secondly, the replacing of the meterstick in a sequence of positions invites a kind of random error. If one repeats the measurement a number of times, different estimates of the length are obtained, the frequency distribution of measurement usually being approximately normal. This kind of random error of measurement is called "noise". It is a reduceable but unavoidable source of error in every measurement. Other sources of error that are of importance here are due to defects in calibration of the measurement scale and to something called "interaction". An example of calibration error may be seen in inexpensive measuring devices where visual inspection reveals that the marks on the scale are not uniformly spaced. Interaction, as a source of error, refers to a phenomenon inherent in all measurement. That is, the very act of measurement causes the thing being measured to change. The measurer interacts, unavoidably, with the object of his attention and thereby obtains a measure which contains a certain amount of error due to the measurement process itself.

Now we recall that a criterion test is used as a device to measure proficiency of an individual pupil with respect to a specified content objective. The measure obtained may be affected by all the sources of error mentioned above. For example, if the questions were 5-alternative multiple choice, there would be an expected "bias" in all observed proficiencies of +.20. Secondly, if the items were, in some sense, of uneven quality, an error akin to poor calibration would be introduced. Third, transient personal and environmental factors introduce "noise"-type errors into the measure; and finally, the pressure and anxiety inherent to some degree in all test taking account for some interaction error. A problem that must be faced, then, is the reduction of error by control of the sources of error. This is done systematically in the following ways.

Criterion-test items are likely to be most accurate when they are of the constructed-response, rather than multiple-choice, type. The process of random selection through computer item-generation is also considerably simplified in the constructed response case. That is not to say that the restriction is essential or necessary, but advisable in most cases to reduce error due to bias.

Secondly, the *a priori* definition of a population of items which are "instructionally" parallel reduces calibration errors. This is accomplished by insuring that all items share common attributes defined by the decision rules which are used to generate the items in a given sample. As much as possible, the items produced will be instructionally equivalent in terms of relevance and importance to the goals of instruction. In this regard, one must also take into account the probabilistic nature of a proficiency measure *vis a vis* the deterministic nature of the illustration given above concerning length. At some sufficiently coarse level of measure, one can unequivocally decide whether or not a distance to be measured is greater or less than the relatively large unit one begins with. However, the presentation of one test item at some "coarse" level never lends itself quite so clearly to a simple decision. One must repeat the measurement, or item, in parallel form to build up confidence in the measure finally obtained. Furthermore, natural subunits are not easy to obtain for a given item. One possible procedure for establishing units and subunits, however, lies in the direction of hierarchical task analysis. One might begin asking questions at a "high" level of complexity, a "coarse" level corresponding to perhaps the use of the meter in length measurement. A few such items should provide a quick estimate of gross proficiency in a broad skill area. Next, hierarchically "lower" subunits of component skills could be measured, building up in this way a measure of each subproficiency. (The tests in Appendix A illustrate roughly the idea involved.) In practice, then, this implies that one would not add or average scores of criterion tests taken at different levels in the hierarchy, just as one would not misalign his columns to add the values of meters to decimeters. Although nothing so neat as decimal notation is implied, the confounding of effects from separable hierarchical levels should be avoided. Adhering to this procedure should also yield better diagnostic information in cases where poor proficiency is measured at some high level by enabling one to measure and examine the component proficiencies separately.

Summing up this view of a CRT as a measuring instrument, we first note that the information contained in a CRT measure is the proportion, rather than the count, of "yes"-decisions made in a sequence of n essentially equivalent replications. The "unit" which accompanies this "numeric" is as essential as stating 6 meters, for example, and not just six in a distance measurement. That is, to say a pupil has a proficiency of .87 is not meaningful by itself. To say that he is 87% proficient on a given class of problems is meaningful for instructional decision-making purposes.

Thus the SCO provides the analog of a unit of measure; the proportion correct is the analog of the numeric. Then, for the same reasons that one does not confound measured information by combining values associated with different levels or units of measure, the score of separate criterion tests are not added together if measuring is to be preserved. Rather the component scores are regarded separately as sources of diagnostic information, to help prescribe further instruction to correct deficiencies measured at higher levels of complexity. This procedure is referred to as a "convergent" test strategy.

2.35 A CRT Performance Model

The definitions formulated so far suggest a number of metaphors that we elaborate now. Our definition of the proficiency parameter, ζ, suggests that we think of the pupil responding to each item in the CRT sample with a fixed probability of correct response. This is consistent with our usage of proficiency in the singular form. It means, from another point of view, that we assume that no "learning" occurs during the time of the test administration which affects the pupil's proficiency. We are not saying that such learning cannot occur, simply that if proficiency changes during the course of the test-taking event, our model will not predict the outcome very well. Again by employing prudent methodology, it is possible to take steps to keep check on the validity of this assumption. Thus we think of the student having one, not many, proficiencies with respect to a well-defined category of problems at any given time.

We also think of the student's responses being independent, that is, we assume that the outcome of any trial is independent of the outcome of every other trial on a CRT. This, of course, amounts to a restriction in the way we generate CRT items, e.g., we must not pyramid problems so that one particular problem holds the key to solving one or more other problems.

Formally, this assumption of independent responses is syntactically defined as follows:

Let n = number of items on the test.

U_{ga} = random variable denoting the $\underline{a}^{th}$ student's response to the g^{th} item (U_{ga} = 0 or 1).

u_{ga} = an observed value of U_{ga}

Then the "local independence" assumption is equivalent to imposing the condition:

$$\text{(2-1) Prob}\left(U_{1*} = u_{1*},\ U_{2*} = u_{2*},\ \ldots, U_n = u_{n*} \,\middle|\, \zeta\right) = \prod_{g=1}^{n} \text{Prob}\left(U_{g*} = u_{g*} \,\middle|\, \zeta\right)$$

The pragmatic implications of local independence extend beyond item writing or selection techniques mentioned above. For a homogeneous proficiency group (i.e., one in which all members have ideally the same proficiency, ζ), local independence says that erroneous responses occur randomly. Therefore, the item intercorrelations calculated using response data gathered from such a group will have an expected value of zero. This property can be used to measure the homogeneity of the group if one can also show that the independence is necessarily a function of uniform proficiency and not purely the result of random errors of measurement. That is, it must be demonstrated that for groups having nonzero true variance in proficiency, the expected item intercorrelations are nonzero and hence, reliability estimates such as KR-20 are nonzero. For example, the data summary shown in Appendix A, page 5, shows for six samples of items independently drawn from three SCO's that KR-20 ranges from .70 to .88 on 5 item tests when computed on the basis of class response data. But when the class is stratified to form relatively homogeneous groups, KR-20 values are sharply reduced and in some instances go negative.

This data does not "prove" that the local independence assumption holds for all CRTs. It does illustrate the kind of test to which one can put the assumptions of the CRT model, however. In the particular case at hand, it indicates that the measurements consisted of something more than random noise. It also suggests the use of KR-20, not as a measure of test reliability, but as a sensor or indicator of group homogeneity. This is, after all, what instructional management is all about: the formation of learning groups whose members are homogeneous with the need for common instructional treatment. One could, for example, use a sliding criterion to advantage. Beginning with a zero error rate criterion one would test for homogeneity within the group of pupils included at each position of the criterion. As wider ranges of scores are included, KR-20 will likely increase. As soon as it passes some set level, say .50, one could back the criterion value of one step and consider all pupils on the low error side of criterion to be a relatively "homogeneous group of masters." A similar procedure beginning with error criterion set at 1 and step-wise decreasing would net a homogeneous group of nonmasters. There could remain one or more distinct intermediate homogeneous proficiency groups. However, preliminary data suggests three groups may be all that can be detected with short tests, of the length required by practical time constraints. In any case, one can thus use the CRT data to classify pupils into relatively homogeneous groups. The members of each group would then be assumed to share common instructional needs for which prescriptions may be provided on the basis of CRT diagnostic evidence.

An illustration of another kind is interesting to note in connection with the local independence assumption. This has to do with the "single-trait" myth. As expressed by Magnusson (1966), for example, this myth says that "An intercorrelation of zero between items means that every item measures something different from every other item." This is, of course, plausible in the norm-referenced frame. However, one could hardly argue, from an instructional viewpoint, that ten randomly selected 2 digit by 2 digit addition problems which do not require regrouping each measures a different trait. Yet it is quite possible that a class can learn to solve such problems with sufficiently uniform proficiency that item intercorrelations do not differ significantly from zero. If the same tests were administered to, say, a mixed group of beginning first graders and graduating sixth graders, the same items would show nearly unit correlations: i.e. all the sixth graders would get them right and all the first graders would likely get them wrong. In the zero-correlation case, the scores are distributed as the binomial with mean ζ. In the unit-correlation case, the distribution is strongly bimodal, consisting of at least two distinct binomial distributions with means $\zeta_{low} \approx 0$ and $\zeta_{high} \approx 1$

The particular point to bear in mind is that these statistics can be used to reveal information about pupils rather than items. Classical test construction treats items as sources of information on certain mental traits. The mental traits are considered to have an a priori existence. The problem, given this context, is to find items with the right properties, properties which ensure that the trait is properly measured. But the criterion test takes the class of problems as the given reality, the objective of instruction, and it attempts to measure the effectiveness that instruction has in changing the pupil's mental attributes, in particular, the proficiency of his problem solving schemes.

This argument leads us to a CRT performance model, i.e., a semantic definition of the item-sampling model. This performance model says, in the light of our previous assumptions that, if the items are uniformly difficult, relevant, and important in the eyes of an individual pupil, then his *CRT performance may be viewed as a sequence of independent Bernoulli trials, each having the same probability of success*, ζ_{ak}.

Hence it follows that if an individual were to be repeatedly given tests consisting of random samples of size *n* drawn from SCO_k, his score distribution would be given by

(2-2) $$f(x_a) = \binom{n}{x_a} \zeta_a^{x_a} (1-\zeta_a)^{n-x_a}$$

where

(2-3) $$x_a = \sum_{g=1}^{n} u_{ga}$$

(2-4) $f(x_a)$ = relative frequency of occurrence of test score x_a

(2-5) $\binom{n}{x_a} = \frac{n!}{x_a!\,(n-x_a)!}$, the binomial coefficient

According to the item-sampling model, each examinee responds to an item *as though* he were tossing a coin with bias ζ_a.

2.4 CRT Statistics

The following are well-known properties of tests built according to the item-sampling model (Lord and Novick, 1968, p.251):

1. The observed test score, x_a, is a sufficient statistic for estimating ζ_a. "Sufficient" means that no information is lost by reducing

the data given in the item-response vector

(2-6) $\underset{\sim}{v} = (u_1, u_2, \ldots, u_n)$

through use of the scoring formula given above in (2-3). Furthermore, if the items are, in fact, parallel then x_a is the minimal sufficient statistic for estimating ζ.

2. Error of measurement is defined syntactically by

(2-7) $\eta_a = x_a - \underline{n}\,\zeta_a$

Since the expected value, $\mathcal{E}(x_a) = n\,\zeta_a$, it follows that the expected error over repeated testing of a given examinee at a given point in time is zero. Pragmatically, this means that a longer test of, say, m·n items, considered as a battery of m parallel tests each of length n, will provide a better true score estimate than a test of only n items provided that one does not make the test so long as to encounter noise due to fatigue.

Clearly, the distribution $f(\eta_a)$ is the same as the distribution of $f(x_a)$ except for a shift in origin. Therefore it follows that the conditional distribution of the error of measurement is not independent of true score since x_a (the observed score) is not independent true score. That is to say, both the error, η, and observed score, x_a, have binomial distributions which are functions of the true score, ζ_a. Thus the criterion-test model differs in this important respect from the classical test model (Gulliksen, 1950) which assumes that errors of measurement are distributed independently of true score.

3. Error variance, the CRT's standard error of measurement, for

individual #a is completely determined by test length and proficiency:

(2-8) $\sigma^2 (n_a) = n \zeta_a (1-\zeta_a)$ (standard error of measurement)

An estimate of error variance, unbiased over repeated item sampling, is derived from the relation

$$\hat{\sigma}^2 = \left(\frac{n}{n-1} \right) \sigma^2$$

$$\therefore \quad \hat{\sigma}^2 = \left(\frac{n}{n-1} \right) \cdot n \left(\frac{x_a}{n} \right) \left(1 - \frac{x_a}{n} \right)$$

(2-9) or $$\hat{\sigma}^2 (n_a) = \frac{x_a (n - x_a)}{n-1}$$

The variance of the error of measurement is a maximum when

$$\frac{\partial (\sigma^2)}{\partial \zeta} = n - 2n\zeta = 0$$

or when

$$\zeta = 1/2$$

2.50 Implications of the CRT Model

2.51 For Item Selection

The item-sampling model described here as the paradigm for CRT construction is one of the simplest of test models. It places no condition on the items except, to preserve score meaning, all items must share at minimum the objective attributes which serve to characterize an SCO. Mixing of items from different SCO's results in the kind of confounding that would occur in any measurement if measures of different kind were combined into a single count.

The SCO not only preserves score measuring, but also defines the scale of measurement. An absolute zero is the least possible proficiency; 1.0 is the maximum. It is easy to see how, if the SCO is not delimited and defined a priori, the scaling falls to an interval level (at best) and why, in that case, one would have to resort to classical item selection procedures for building a measurement scale. With an unrestricted item-population, such as one consisting of items "like those" in a given sample, there is no evident limit to how well or poorly students might do on replicated tests. Presumably one could find sufficiently easy items that any pupil could do at least a few of these correctly and by the same token, the existence of sufficiently hard items would prevent anyone from doing all items well. Thus an absolute zero point could not be assumed to exist under such assumptions. Furthermore, absolute values of observed scores would be a function of the mean item-difficulty in the selected sample. By biasing the item selection process to favor items of a given difficulty, it is possible in such cases to build tests having some predetermined class mean. Thus the absolute value of the observed score would not be meaningful. Only the ranks, and possibly the differences between ranks, would preserve their meaning when the item population is not well-defined.

By contrast, pupils not familiar with the problem solving skills involved in the items oound on a given CRT can and will show a true zero proficiency. Since the SCO is sharply defined, one cannot search about endlessly in an infinite pool in search of items which discriminate at arbitrarily low ability levels. There simply aren't any such items available within the defined population. The corresponding argument holds at the high proficiency end as well; the scale of the CRT is limited by an individual's ability to do all problems in the defined population.

The implications of this "scale-definition" property of the SCO can be seen in the error expression, (2-8). It says that measurement error arising from "noise", or random variations in observed score, is a function only of test length, n, and pupil proficiency, ζ_{ak} and is not dependent on any item parameters. Thus, from the above argument and equation (2.8) the first implication to note is that the use of classical item selection procedures can in no way reduce or modify the CRT error of measurement. These classical procedures are functions of item parameters defined for norm groups of persons and therefore involve variables that are not part of an item-sampling model. Thus the only critical consideration regarding item selection for a CRT is the random selection of items from among a well-defined population.

2.52 For Test Reliability

A second implication concerns reliability of measures. Reliability is used here to mean a measure of the degree to which replicated measures agree. The reliability in this sense, is a function of error variance. The smaller the variance, the higher the reliability. Equation 2.8 says that error variance for a CRT is a function only of n and ζ. The error variance, $\sigma^2(n)$, indicates the amount by which observed measures, obtained through replicate testing on an individual, deviate from the sample mean. Each sample mean in turn, estimates the item population mean for the individual. Since the CRT should estimate this population value reliably, a measure of a CRT's "reliability" is the standard error of measurement, i.e., the standard deviation of the random sampling distribution of sample means given by

$$\text{(2-10)} \qquad \text{S.E.M.} = \sqrt{\frac{\zeta\,(1 - \zeta)}{n}}$$

This too is independent of item parameters and dependent only on n and ζ.

Again by contrast, NRT reliabilities are functions of item parameters. For example, a lower limit to NRT reliability is given by coefficient alpha. (Lord and Novick, 1968, p.331):

$$\text{(2-11)} \qquad \alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{g=1}^{n} p_g q_g}{\sum_{g=1}^{n} \sigma_g \rho_g}\right)$$

where n = number of items

p_g = item difficulty estimate

$q_g = 1-p_g$

ρ_{gx} = item-score correlation

Clearly α is a function of such item parameters as (1) item variance and (2) item-score correlation, which in turn is a function of item intercorrelations. Furthermore, if the NRT score, x, is desired to correlate closely with a predicted variable (external criterion), Y, one finds that the NRT score-criterion correlations (called test "validity") is also a function of item-criterion correlation, item variance, and item intercorrelations.

The basic reason why classical reliability formulas are complex functions of item parameters and CRT reliability is a simple function of pupil proficiency and test length lies in the distinct notions of reliability involved. The NRT needs not only reliable estimates of scores, but also maximum dispersions between scores in order to achieve replicable rankings. Differences in true scores must be magnified, so to speak, so that errors of measurement will not cause many inversions in rank to occur in test replications. The underlying assumption in the NRT is that one expects to find true differences in rank in any sample of persons. The psychometrician operates on the primary assumption that such differences are due to normally distributed differences in underlying traits, mental traits whose measure is his chief concern.

The CRT, as an instructor's tool, reflects the view of Carroll (1963) and Bloom (1968) that differences in native ability can be compensated for by individualizing the pace and method of instruction. Thus it is conceivable that among certain samples of persons, true or significant differences in proficiency may disappear as a result of instruction. Uniform achievement is the ideal expected. Thus one can be content to obtain replicable measures of the common proficiency even though such a test could obviously fail to meet the additional requirement of ranking reliability which is characteristic of the NRT.

Perhaps most important to note is that, for the instructor, the instructional goals described by certain SCO's are the given quantities, and the mental skills which account for proficiency are the learning variables. In other words, items are fixed and mental abilities are to be developed by instruction. In the NRT case, mental abilities of interest are the given aspects of reality; the task is to find items (the variables) which involve the use of an existing ability and thus measure the degree of its presence or absence. The latter is complicated by the fact that little is known of item/trait relationships. Thus one must rely on complicated inferences drawn from item data obtained in pilot testing with representative pupil groups.

2.60 Comparison of the CRT Model with Other Test Models

At the semantic level, the essential point of difference between the CRT and other test models is that the CRT measures proficiency whereas virtually all existing mental test models are designed to measure certain assumed mental, or latent traits. Whereas it is reasonable to assume normal distribution of the latent traits, we have indicated a point of agreement with Bloom (1968) that such an assumption is not warranted with respect to proficiencies following content-specific instructional treatment.

This distinction suggests the direction in which to look for points of difference between the CRT and other models at the syntactic level. We have already noted syntactic differences between the CRT and classical models. Even the item-sampling model discussed by Lord (1968); Cronbach, Schonemann and McKie (1965); and by Schoemaker (1966) are essentially latent trait models. These require, not just item sampling, but matrix sampling across both items and populations in order to obtain generalizable estimates of the means and variance for population latent traits.

Furthermore, none of these item-sampling models impose the essential restriction involved in the CRT model expressed by the definition of the SCO.

In a general way, one might say that the CRT is subsumed as a form of the quantal response model, a class to which the normal ogive and logistic model also belong. The syntactic points of agreement are extensive (cf. Lord and Novick, 1968, p.420-435) between the CRT model and other forms of the quantal response model. These may be summarized as follows:

Let $\underset{\sim}{v} = (u_1, u_2, \ldots, u_n)$ be the observed response vector of an individual to an n-item test. Let V be a random variable denoting any possible response pattern. Then the general quantal response model is given by

$$\text{Prob}\ (\underset{\sim}{V} = \underset{\sim}{v} \mid \theta)$$

The logistic and normal-ogive treat θ as an unbounded vector of latent traits which underlie test performance; the CRT treats θ $(=\zeta)$ as a bounded scalar, called proficiency. In spite of the syntactic similarity between the various forms of the quantal response model, the logistic and normal-ogive lead to test design based on considerations of item parameters and characteristic functions; the CRT does not.

Three other models which are similar to the CRT in some respects are those associated with Guttman perfect scaling, Rasch's Poisson

Process model, and Lord's strong true-score theory based on the binomial error model. Again the essential difference arises in the interpretation of the measure. The CRT is designed to be generalizable for an individual across a population of items; most other mental test measures are intended to be generalizable across a population of items and/or a population of persons.

The Guttman scale treats item difficulties according to the "latent distance model": (Lord and Novick, 1968, p.547)

$$\Pi_i \ (\theta) = \begin{cases} 0 \text{ if } \theta \leq c: \\ 1 \text{ if } \theta > c: \end{cases}$$

If the items are ordered so that $c_1 < c_2 < \cdots < c_n$, each item defines the cutoff point for a distinct latent class. If CRT's are hierarchically ordered, one similarly obtains an ordering of proficiencies which define content hierarchies:

$$\zeta \ (SCO_k) \approx \begin{cases} 0 \text{ for } k \leq c: \\ 1 \text{ for } k > c: \end{cases}$$

The parameters ζ and θ are, however, semantically distinct. ζ is essentially equivalent to a true score concept while θ lies one level of inference beyond that, at the latent trait level.

Rasch's "Poisson Process Model" goes beyond the CRT item-sampling model to attempt to account for proficiency in terms of separable items and ability parameters. The model is interesting in its implications for person-free and item-free test calibration but of negligible use in the instructional management functions of concern in this discussion since, again, item characteristics become involved.

Finally, Lord's Binomial Error Item-Sampling model is closely related to the CRT syntactically. The formulation is identical up to a point where Lord attempts to calculate regression of true score on observed score. This leads to an empirical Bayes estimation problem which is concerned with generalizations over person-populations rather than over item populations.

In short, the CRT model is a subform of many of these models but distinct in the use of a well-defined item population. This is the essential restriction which permits effective application of the model to a class of measurement problems which are of interest to the instructional manager, namely the assessment, at frequent intervals, of rapidly changing proficiencies.

CHAPTER III

MANAGEMENT SUPPORT SYSTEMS: THE DECISION COMPONENT

3.0 Introduction

The management of instructional systems basically involves a comparison of the actual learning product with the intended product, at selected points in time, and then making intelligent decisions in regard to subsequent instructional processes on the basis of the comparative data. The subsystem which assists teachers or pupils to perform this comparison and decision function using data generated by the Information Component is called the Decision Component. It is the purpose of this chapter to discuss the advanced design of a particular Decision Component for instructional management derived from the first-order CRT model discussed in Chapter II.

Advanced design of a system component is necessarily somewhat tentative due to the fact that not all information needed is available at the outset. The procedure one follows in the face of incomplete information is essentially a strategy of systematic trial and revision. The advanced design phase takes into account all the available information together with estimates of the effect of other relevant factors which are known only imprecisely at the time. The system is then tried out to gather performance data to be used in subsequent steps of its development. Chapter III, therefore, first deals briefly with the broad range of considerations that affect instructional management system design, then in detail with more limited but more precisely specifiable problems encountered daily in the classroom. In this latter category are included especially (1) the specification of curriculum design requirements and acceptance sampling requirements, (2) the selection of efficient sampling plans for diagnosis and prescriptions, and (3) the decision rules to be applied to the data output from the Information Component in order to form homogeneous instructional groups or to recommend effective prescriptions.

3.01 The Role of Computer and Teacher in Semi-Automated Instructional Systems

One of the first questions usually raised in connection with proposals for automating any part of the instructional process concerns the roles of the teacher and the machine. Since these roles together with the purpose for automation relate particularly to the form the Decision Component takes, we first note in broad outline the function of the Decision Component before taking up the formal details of its design.

The purpose of the envisioned semi-automated decision system is to produce a more humane and a more efficient instructional process than is customarily found in existing, nonautomated systems. In spite of the lip-service given to the child's individual needs when planning instruction, existing instructional systems are more notable for the extent to which such needs are ignored rather than effectively utilized to assist the learning process. Whether it is on the basis of Rice (1893) discussing the schools of 1890 or Koerner (1963) describing the schools of 1960, there is good reason to conclude that our educational system has a tendency to favor administrative utilities at the psychic expense of the learner.

Part of the reason which accounts for this behavior lies in the problems of processing the amount of information that is needed to take individual differences into account when planning instruction. Part of the problems also lie in the inefficiency of classical information and decision systems, which were originally conceived (e.g. by Binet) not with instructional management in mind but for the purpose of identifying mental deviates. It is not altogether surprising that these measurement systems, which employ a unit of measure based on relative deviation from a norm, have not been found to be particularly useful to the instructional manager, whose information needs are of a much more particular and absolute nature. Part of the problem also lies in the fact that limits of time and human endurance are typically exceeded by existing inefficient instructional systems. This prevalent condition constrains teachers and administrators alike to favor the learner's behavioral conformity in the name of discipline at the expense of developing disciplined creativity and the capacity for independent learning. Administrative constraints are therefore imposed to lock children into learning rituals in which feedback rewards the conforming individual, be he teacher or pupil.

Given this value system, it is clear that such individualization of instruction as does occur operates within a severely restricted framework. One source of these restrictions lies in the fact that the system imposes rites of passage on the learner which are a function of chronological age and not of intelligence or ability to learn. These rites are maintained by powerful social structures that, as Koerner (1963) and others have noted, cause the educational system to react to threats of change in a protean manner, constant only in its determination to prevent substantive change. Such stability is evidence not of malicious establishmentarianism, but more likely, of experience with alternatives which when tried in the past have shown only a disconcerting tendency toward failure. The principal cause of system failure can often be traced to the lack of adequate management techniques to control individualized programs of instruction.

Thus it is of little use simply to provide more diagnostic information to an already overburdened teacher. Previous experience with automated management systems indicate that the teacher will not have the time and energy effectively to utilize additional information by itself, given existing instructional systems. Rather it is necessary that the instructional system and the management (=information/decision) system develop together in a coordinated fashion. Since each depends for its effective functioning on the other, it is necessary that the development of improved curricula evolve through systematic trial and revision. This has been described (DeVault and

Kriewall, 1968) as a dynamic curriculum model.

It is not sufficient to expect that a management system will work once an adequate logic has been developed; however, an adequate design may reduce excessive concerns for control and conformity so that opportunities for greater individuality of learning styles will both be tolerated and encouraged. In this way, the evolution of a more humane and intelligent system of education may be made possible by making use of partially automated, labor-saving management systems.

From this it is clear that the role of the machine is not to take over in any sense the legitimate instructional functions of the teacher. Rather, as we shall see, the machine is used to manipulate data gathered from the Information Component and to make certain decisions, according to rules which conform to the instructional intent of the teacher. By also keeping necessary records of pupil performance in the machine, the teacher is freed to do that which the human component of the system is best designed to do, to make intelligent instructional decisions regarding available alternatives in the light of both the data produced by the machine and the judgmental data that is effectively processed only by humans. In short, the computer acts as a quality control inspector but the teacher determines both the standards of quality and, together with the individual pupil, the instructional process by which the product of desired quality is to be obtained.

3.1 Process vs. Quality Control

The development of a systems approach to instruction, at a logical level of analysis, begins with a consideration of what is meant by design and acceptance requirements. To establish the context in which design and acceptance requirements have meaning, we review the essentials of the systems approach to instruction with particular emphasis on these concepts.

An instructional system may be described in terms of its input, process, and output. The general model depicted in Figure 2, Appendix G, implies that a total instructional process may be regarded as being composed of certain unit processes, or "packages", as they have come to be called in the recent literature. We shall consider an instructional unit to be associated with one, or possibly a few, specified content objectives.

The input to each process unit is one or more pupils having entering proficiencies measured by CRT pre-tests. The output of a successful instructional process is a pupil who has developed certain additional proficiencies. An instructional step is completed as the output from one unit process becomes the input for another.

The potential for new behavior learned by the individual is one of two primary products of an instructional process. The other product is information concerning the operation of the system. This information is needed to modify system operation in such a way as to improve the quality of its learning output (cf. DeVault and Kriewall, 1969, Chapter 3). In order to utilize such information effectively, however, one must have knowledge of what was intended or designed into the product when the instructional process was constructed;

also, one must have techniques for determining the degree to which the product conforms to the design requirements.

This view of instruction as a process and proficiency as a product, although oversimplified, enables one to discuss the basic principles of process control and quality control in educational terms. Quality control for the purposes of this discussion, involves the assessment of pupil proficiency before and after given instructional processes; process control involves the manipulation of the instructional control variables so that improved quality is obtained. The basic method of process control consists of comparing the measure of specified characteristics observed in randomly selected samples of the product with standards established at the time that the instructional process was designed. If the proportion of pupils submitting substandard learning products for inspection lies beyond established instructional control limits, steps must be taken to improve the instructional process. Even when nominal requirements are being met, small perturbations in process control variables may be experimentally induced in order to explore methods of producing a higher quality of learning product.

It should be noted that process control differs from quality control in that the latter is a method of deciding, in educational systems, whether or not the individual learner shall go on to other objectives. Process control, on the other hand, is a method of insuring that an adequate proportion of learners will meet acceptable quality requirements upon exiting a given instructional unit. When, for example, an instructional package is said to have a 90/80 design criterion, the first value refers to the process control limit: 90% of all pupils entering the instructional process are intended to meet acceptable quality requirements upon completion of the unit; the second value refers to the quality control acceptance limit: the minimum acceptable proficiency that an individual must achieve is 80%. The latter value establishes a criterion for making decisions regarding individual pupils; the former establishes a criterion for controlling the instructional process.

Thus system output can be characterized in terms of its quantity and quality. Translated into instructional terms, this means, on the one hand, it is desirable that the largest possible proportion of pupils meet acceptable standards of quality and, on the other, that the minimum acceptable quality limit be as high as feasible.

It is evident that a conflict arises between the desire for both maximum quality and quantity. With a given instructional process, the number of students who pass acceptance requirements increases if the quality, or product design standard, is lowered. The higher the standard is set, the more likely the product is to be rejected. Thus maximizing the output in a feasible way requires that one consider system constraints and conflicts which arise in the process of allocating available resources to meet the identifiable needs of individual learners. Time is an important constraint. Money needed to supply materials and services is another. Available human and material resources place limits on what one can expect children to learn successfully. Finally, the availability of diagnostic and prescriptive information, together with the cost of obtaining it, determines in large measure the degree to which immediate learning needs of individual pupils can be detected. This indicates the need for new competencies in educational workers; a need for instructional analysts who can devise techniques for maximizing expected learning output by applying systematic decision-making to the consideration of values contingent upon available alternatives.

The values of concern here deal with the quantity and quality of instructional output; the decision-procedures are to be derived from the CRT model, subject to the constraints listed above.

3.2 Specification of Instructional Design Requirements

Design requirements for an instructional unit specify the explicit characteristics that the learning product is supposed to have. Design requirements are therefore analogous to such notions as instructional objectives, learning objectives, behavioral objectives, etc. The main distinction between the idea of design requirements and the various kinds of objectives mentioned lies in the detailed specification of minimum expected performance levels with respect to an SCO. Rather than say, for example, that a child will be able to discriminate between like and unlike pairs of words, a design requirement specifies the minimal proficiency that is acceptable at a given point in time on a given SCO. An illustrative design requirement for an instructional unit might appear as follows:

TABLE 3-1

Specification of Design Requirements

Specified Content Objective	Expected Proficiency	Available Time	Proportion of Pupils Attaining Mastery
i	ζ_1	$t_1 \pm a_1$	P_1
$j \supset i$	ζ_2	$t_2 \pm a_2$	P_2
$k \supset (j \cap i)$	ζ_3	$t_3 \pm a_3$	P_3

The first column indicates the sequence of content objectives and their logical hierarchy as determined by task analysis. (e.g. SCO j requires SCO i as a prerequisite; SCO k requires both j and i). The second column indicates the design requirements in terms of minimal acceptable proficiencies; the third column indicates the tolerances in time available for instruction; and the fourth column indicates process control limits: the proportion of entering pupils who successfully meet acceptance requirements upon completion of the unit.

Design requirements for an instructional system are formulated as a part of curriculum development. It should be remembered in this regard that instructional production, in contrast to industrial production is spiral rather than cyclic in its organization. The evaluation of a learner's proficiency should, therefore, be relative to his state of development. If he is at the lower reaches of the spiral, one might be quite pleased with a 50% or 75% proficiency in some class of problems. The main point of instruction at such a level could be impressional rather than functional: the teacher may simply wish to expose the child to new ideas and develop minimal beginning proficiencies. Later, the instruction may return again to the topic, extending it, relating it to other ideas, while at the same time raising the level of functional proficiency to perhaps 80 or 90%.

From these considerations it should be clear that an adequate curriculum design must utilize a more flexible concept of mastery than that which has been inherited from the disciplinary era of the last century. Mastery must be considered as a local or relative value that is a function of the child's development in a well-designed and articulated curriculum. At the other extreme, the sentimental notions associated with such school movements as the expressionism of the 1920's and the incidentalism of the early 1930's are inadequate because they tend to ignore the need for specifying design requirements, in particular the expected levels of achievement.

Inflexible mastery levels should therefore by replaced wherever they exist by relative levels that fit both the learner's needs and the rationale of a spiral curriculum; but, on the other hand, objective requirements should be maintained where possible in lieu of subjective or sentimental expectations.

That these are legitimate present day concerns is evident both in the perspective of our educational heritage and in the practice found in schools of the day. For example, there is the concept of unconditionally successful instruction which enjoys newly found currency in some schools. It is characterized by a design requirement which sets a rigid 100% level of proficiency as the minimum acceptable standard. Primarily, it exemplifies the fact that inflexible mastery standards still intrude on present day instructional planning. At the opposite extreme, one finds the growing popularity of an inquiry approach to learning translated into instructional arrangements for which it may happen that no design requirements at all are specified. Without adequate control, very low proficiencies could be the unintended product of such instructional patterns. Between these extremes lies the set of admissable design requirements. The task of the instructional designer is to find and specify reasonable performance requirements for each instructional segment, or package.

It is usually not a reasonable goal of instruction to strive for 100% levels of proficiency. The cost of programs with so-called high standards very likely would be astronomical while the utility to the individual or society, gained through marginal improvement at the extreme level of proficiency, would not justify the cost involved as a general practice. So while it is attractive to speak of the pursuit of excellence, unconditionally successful instruction, and other utopian goals, it is perhaps more realistic to follow the maxim that it is better to lower one's sights in order to hit the bullseye than to aim high and consistently miss the mark. Systems design requirements, in other words, should aim at rational decision-making that yields optima as contrasted with utopian decision-making that seeks bonanza without due regard for the possibility of disaster.

The writing of instructional design requirements calls for the highest levels of professional competence. One must be familiar with the logical development of subject matter as well as with the cognitive, affective, and psychomotor development of children. Obviously, the state of the art is presently such that this phase of curriculum construction is indeed very much art and very little science.

Having noted the need for specifying design requirements and a philosophy to guide the writer of such specifications, we turn our attention to the complementary problem of enforcement: what evidence, gathered by what means is needed as proof that the learning product coming from an instructional process is acceptable?

3.3 Specification of Acceptance Requirements

Just as design requirements specify the demands one makes on the performance of an instructional system, acceptance requirements are needed to specify the evidence that will be considered sufficient to accept the product. Acceptance requirements therefore are composed of two kinds of specifications: (1) The methods to be used in testing the product and (2) the sampling procedure by which one determines whether or not minimal requirements have been met. The method of testing advocated here involves the use of criterion tests associated with the particular SCO's of instructional interest; the sampling procedure is the principal topic of the next section.

In a general sense, the acceptance requirements for an instructional process can be specified in a manner similar to that used in industrial acceptance sampling procedures. However, there are important differences which we note at this point. Following inspection of a sample from a batch produced by some industrial process, the important decision is whether or not to accept the batch for delivery to the consumer or reject it. Rejected batches may be scrapped or they may be screened of all defective items before delivery to the consumer or the defective items may be reworked until they meet standard specifications. Following an instructional process, one can also select samples of items from batches, in a certain sense, and make a decision to accept or not accept the entire lot. The questions that need to be clarified here concern first, what constitutes an item of learning product; secondly, what is to be considered as the analog of the batch or lot; and thirdly, what is the nature of decisions to be made concerning lots that are accepted and those that are not accepted.

The items of instructional production are essentially the "problem solving skills" that the learner has developed during the instructional process. The analog of the lot or batch is the set of skills required to perform successfully on problems selected from a given SCO. The sample of items submitted for inspection is determined by the random sample of items selected from the specified content objective, i.e. the pool of problem-situations defined by the SCO. Since we have restricted our CRT model to the consideration of binary items only, the response to each problem is judged to be either right or wrong. If the proportion of errors is less than the criterion stated in the design specifications, a decision is made to accept the set of problem solving skills associated with the entire population of problems in the SCO without further observation. If the errors exceed the criterion, an instructional prescription will need to take this fact into account.

The formulation of design requirements, as noted above, involves the problem of looking ahead in time to forecast what knowledge and proficiency an individual needs for continued educational success. These needs must be carefully adjusted within the limits of what is possible both in terms of the child's development and available school resources.

Formulation of acceptance requirements involves a much different set of problems. The foremost problem facing the writer of acceptance requirements is economy of time and money. In rough outline the situation is as follows:

The pupil who completes an instructional unit may or may not have attained the minimal proficiency with respect to the SCO specified in the design requirements. Therefore, it is necessary to test his proficiency on the SCO item population to determine whether or not he has developed acceptable proficiency. The amount of time and the expense required for such testing is a system cost. Acceptance requirements must be written as to minimize this cost.

The important parameter to consider in regard to cost is the number of test questions, $\underline{n}$, which we refer to as test length. The reduction of $\underline{n}$ decreases cost, of course, but at the same time it increases the probability of making erroneous acceptance decisions. Since the error of measurement increases as $\underline{n}$ decreases, the probability of misclassifying both "masters" and "nonmasters" also increases. Misclassification represents another kind of subjective cost, or disutility, to the system. The one cost, therefore, must be balanced against the other when writing acceptance requirements.

There is also another important utility to be considered in connection with reducing test length, $\underline{n}$. Suppose, for simplicity of argument, that all the questions that might be asked of a student in a semester are ordered linearly in time. The SCO, as described above, represents the organizational technique used to group these questions into classes analogous to industrial inspection lots. Suppose there are $\underline{k}$ such classes of questions, or SCO's. Suppose further that one samples n_k items from each SCO_k. If $\bar{t}$ represents the mean time required for a pupil to take a CRT, then the total required testing time to inspect all $\underline{k}$ SCO's is the product, $\underline{kt}$. If total available time for testing in a semester is denoted by T, then it is necessary that

3-1: $k\bar{t} \leq T.$

However, the usual situation is that condition (3-1) is violated, i.e. the required time for testing exceeds the available time. One of two remedies are possible if T is fixed: either reduce the number of objectives, k, on which management data is gathered or decrease the average time required to administer a test. The latter, of course, is the more desirable alternative and this is accomplished by reducing the test length, n_k, to the minimum required for efficiency.

In general, it can be seen from the condition (3-1) that minimizing n_k enables one to maximize the number of objectives, k, that are brought under effective control of the management system. Since techniques of diagnosis and prescription depend for their precision on the number of SCO's for which data is available, one has additional reason for minimizing test length.

Order-of-magnitude figures indicate that conventional classroom procedures permit time for approximately 1000 test items to be asked during one semester, per subject area. If 10 items are needed for sufficiently reliable classification decisions, it follows that only about 1000/10 = 100 objectives can be "managed", since that is all the data available to the decision component. If it were possible to halve the average needed test length, one could bring twice as many learning objectives under explicit management control. Among other things, such estimates indicate the futility, from a manager's standpoint, of mindless decomposition of learning objectives to finer and finer levels of specificity. There is need only to specify curriculum design requirements for those objectives one intends to sample and enforce quality standards on. This conclusion holds independently of the type of management system employed. The intent of partial automation, of course, is to increase the number of instructional objectives that are effectively managed as compared with the 100 or so per subject per semester that an unaided teacher can manage efficiently.

Thus it can be seen that the problem of specifying proficiency acceptance requirements is to a large extent one of considering practical economics and mathematical probabilities whereas the specification of instructional design requirements involves a much broader spectrum of considerations. However, the development of better acceptance requirements should help to ease the task of preparing better instructional design requirements; for as we consider ways of determining whether we have produced what in fact we wanted to produce, we are likely to gain a clearer picture of what was really wanted in the first place.

3.4 Principles of Inspection Sampling of Learning Products

Once given a set of design specifications, the problem of specifying acceptance requirements reduces to the selection of an adequate but economical sampling plan which efficiently classifies pupils into homogeneous instructional groups. A method of selecting efficient sampling plans for diagnosis and prescription derives from the Criterion-Test theory of Chapter II. This derivation will be completed in two steps as follows.

First a numerical example will be employed (1) to motivate and illustrate the definitions of certain sampling concepts and (2) to show how the quantal response and item-sampling models are actually models of two kinds of probability sample spaces. In the quantal-response case, a "point" or elementary outcome in the sample space is semantically representative of an item-response pattern whereas in the item-sampling case, an elementary outcome is simply the test score.

Although the sufficiency of the test score statistic, x_a, has been established under the assumptions of the item-sampling model, the numerical example is intended to demonstrate the practical implications as well as the plausibility of the theoretical argument for ignoring item-response patterns in favor of looking only at CRT test-scores for diagnostic purposes.

The second step leading to the selection of efficient sampling plans for diagnosis and prescription involves going from the particular case of the illustrative example to the general case derived from item-sampling theory. At the outset, two analytical tools are adapted from familiar sampling theory. The lattice diagram is introduced as a visual representation of a CRT sample space and the cumulative binomial distribution is employed to define the probability measure on this point-set. Next, the "operating characteristic" for a CRT is derived and shown to function analytically like an item characteristic curve in the sense that test length determines the CRT discrimination (like an item-"β")* while criterion-selection determines the point on the proficiency continuum at which the CRT's maximum discrimination occurs (in the manner of an item "X50")*. The theoretical development culminates in a discussion of practical acceptance requirements related to the following type of instructional management problem.

---

*cf. Baker, F. Test Analysis Package, The University of Wisconsin, 1966.

A child's performance on a CRT is a sample of problem solving skills which are being submitted for inspection. Acceptance requirements specify how large this sample must be and, in some applications, what criterion for acceptance will be used in order to provide a specified level of protection against possible errors of classification. This specification of a test procedure defines a _sampling plan_.

The basis for selecting a sampling plan, S, that is as good or better than alternative possible plans, S', is an optimization problem that lends itself to a decision-theoretic treatment, (e.g. Wetherill, 1966). Since such selection procedures involve the assignment of utilities or costs for which little information is available at this time, we shall only outline various possible strategies for selecting optimum sampling plans but analyze and predict the protection given by some _ad hoc_ sampling plans that appear to be most promising for instructional management systems, given the present state of costs and technology. These plans will be called (1) the Single Sampling Plan (SSP), (2) the Curtailed Single Sampling Plan (CSSP), and (3) the Sequential Sampling Plan, labeled the "SPRT" after Wald's (1947) Sequential Probability Ratio Test.

---

*cf. Baker, F. _Test Analysis Package_. The University of Wisconsin, 1966.

3.41 Test Outcomes as Points in a Probabilistic Sample Space

In order to select acceptance requirements which have a predictable level of efficiency for correctly indicating whether or not a learning-product meets design specifications, it is useful to consider inspection sampling as an "experiment" involving random phenomena. From this probabilistic viewpoint, the quantal response model and the item-sampling model discussed in Chapter II are, in effect, models of this imagined experiment. Which model applies in a given case will be shown to depend on the form the data is in when the acceptance standards are applied. The essential ideas are illustrated in the following simple numerical illustration.

Suppose that pupil _a_ has completed an instructional package on a certain class of long division problems (e.g. Test #19, Appendix B). Let us assume that the design requirements for the package state that 50% of all students completing 2 ± 0.5 instructional periods on this specified content objective will develop a proficiency of at least 0.67. Table 3-2 summarizes these design requirements.

TABLE 3-2

Illustrative Design Requirements

SCO	ζ'_1	Available Time	Proportion Successful
Long Division: SCO #19	0.33	2 ± 0.5 pds.	0.50

Now suppose we specify acceptance requirements for this instructional process by selecting a sampling plan with $n_k = 3$ and an error criterion c = 2. The acceptance requirements for the proposed sampling plan are summarized in Table 3-3:

TABLE 3-3

Illustrative Acceptance Requirements

Sampling Plan	n	Error crit. c	Decision Rules
SSP	3	2	1.If $n-x_a < c$, accept H_0 2.If $n-x_a \geq c$, accept H_1

The symbols H_0 and H_1 denote respectively that design requirements are met by the learning product (H_0) or that design requirements are not met (H_1).

A number of questions are explored below which illustrate two basic problems of instructional acceptance sampling: first, is the length of the test adequate; secondly, is the criterion set at an advisable level?

It will be convenient to define a nominal student as one whose proficiency meets the minimal acceptable design requirement. We shall denote by α the probability that a nominal student is misclassified (i.e. the probability of an error of Type I occurring). It should be evident that α represents the largest probability for an error of Type I since if any student has a proficiency higher than the nominal limit, his chances of making c or more errors will be less than those of the nominal student, hence the probability of his misclassification will be less.

In the sense defined earlier, α is a measure of the efficiency of the selected sampling plan. In order to calculate this efficiency, it is necessary first to compute the probability of the possible outcomes for a given examinee.

We imagine that each pupil *a* takes the three item test, resulting in a test outcome that can be described in terms of an *item-response vector*:* $\underset{\sim}{v}_a = (u_1,u_2,u_3)$. To be explicit, all the possible test outcomes are included in the following eight patterns:

$$\underset{\sim}{v}_0 = (0,0,0);$$
$$\underset{\sim}{v}_1 = (0,0,1);$$
$$\underset{\sim}{v}_2 = (0,1,0);$$
$$\underset{\sim}{v}_3 = (0,1,1);$$
$$\underset{\sim}{v}_4 = (1,0,0);$$
$$\underset{\sim}{v}_5 = (1,0,1);$$
$$\underset{\sim}{v}_6 = (1,1,0);$$
$$\underset{\sim}{v}_7 = (1,1,1);$$

A probabilistic model can now be constructed by considering each vector $\underset{\sim}{v}_{ia}$ as an elementary outcome in a probability space, Ω.

---

* The notation adopted throughout this paper conforms generally with the recommendations of the Committee of Presidents of Statistical Societies as published in the *American Statistician* of June, 1965. A list of the notational conventions used is given in Appendix C.

A probability space is a set, in the mathematical sense, on which a measure, called the probability, is defined. In the case of the illustration here, the probability space, Ω, is defined as follows:

$$\Omega = \left\{ \underset{\sim}{V} \,\middle|\, \underset{\sim}{V} = \underset{\sim}{v}_i \quad \text{for} \quad i = 1, 2, \ldots, 2^{n_k} - 1 \right\}$$

where $n_k = 3$, the number of items in the hypothetical test.

The probability measure is defined by using either the quantal response model or the item-sampling model for the CRT. We shall first show an application of the quantal response model primarily for the purpose of demonstrating the following facts. First, there is no information contained in the detailed response pattern $\underset{\sim}{v}_{ia}$ that is lost in data reduction by the use of a scoring formula; and secondly, that the appropriate model to use on the *reduced* data is the item-sampling model.

The quantal response model permits us to calculate the probability of each elementary test outcome occurring, given the proficiency, ζ. Recall that proficiency, ζ, is interpreted as the probability of a correct response to a randomly selected item; that $\zeta' = 1 - \zeta$ denotes the probability of an incorrect response, or the "error-rate"; and that the CRT performance model implies that we may idealize a pupil's performance on the CRT as a sequence of n_k Bernoulli trials, each trial having a probability of correct response, ζ.

The quantal response model is Prob $(\underset{\sim}{V} = \underset{\sim}{v} \,|\, \zeta)$. What we need to do, then, is calculate the probability for a random event $\underset{\sim}{V}$

occurring given ζ. Since each item response is an independent trial, the product-rule of probability theory applies: the probability of two independent events occurring in succession is the product of their respective probabilities.

Thus, for example, applying the product-rule to $\underset{\sim}{V} = \underset{\sim}{v}_0$ (=0,0,0) we have

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_0 \mid \zeta) = (1 - \zeta)(1 - \zeta)(1 - \zeta) = (1 - \zeta)^3;$$

since each item-response observed in the pattern v_0 is "wrong". Similarly

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_1 \mid \zeta) = (1 - \zeta)(1 - \zeta)\zeta = \zeta.(1 - \zeta)^2;$$

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_2 \mid \zeta) = (1 - \zeta).\,\zeta\,.\,(1 - \zeta) = \zeta.\,(1 - \zeta)^2;$$

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_3 \mid \zeta) = (1 - \zeta).\,\zeta\,.\zeta = \zeta^2\ (1 - \zeta)\ ;$$

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_4 \mid \zeta) = \zeta.\ (1 - \zeta)\ .\ (1 - \zeta) = \zeta\ (1 - \zeta)^2;$$

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_5 \mid \zeta) = \zeta(1 - \zeta)\zeta = \zeta^2\ (1 - \zeta)\ ;$$

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_6 \mid \zeta) = \zeta.\ \zeta.\ (1 - \zeta) = \zeta^2\ (1 - \zeta);$$

$$\text{Prob } (\underset{\sim}{V} = \underset{\sim}{v}_7 \mid \zeta) = \zeta.\ \zeta.\ \zeta = \zeta^3\ ;$$

The pattern of probabilities becomes clearer if we group test outcomes together which have the same likelihood of occurrence. Table 3-4 summarizes the results in this way:

TABLE 3-4

Probabilities associated with Response Patterns and Test Scores

	Probability of Observed Outcome			
	ζ^3	$\zeta^2\ (1 - \zeta)$	$\zeta\ (1 - \zeta)^2$	$(1 - \zeta)$
Test Outcome:	v_7=(1,1,1)	v_3=(0,0,1) v_5=(0,1,0) v_6=(1,0,0)	v_1=(0,0,1) v_2=(0,1,0) v_3=(1,0,0)	v=(0,0,0)
Score: x_a	3	2	1	0

Given the probabilities shown in Figure 3-4, we can next calculate the efficiency of the sampling plan, selected for these illustrative purposes, as follows. From the design specifications given in Table 3-1, we see that, for a nominal student, $\zeta_1 = .67$. Hence the probability that this nominal student will exceed the error criterion limit and therefore be incorrectly classified as a nonmaster is equal to the probability that at least one of the outcomes v_o, v_1, v_2, or v_4, will occur. Since these are assumed to be mutually exclusive events, the sum-rule of probability theory applies: the probability that at least one of a set of mutually exclusive outcomes will occur is the sum of their respective probabilites of occurrence.

Thus

$$\alpha = 3(.33)^2(.67) + (.33)^3$$

or

$$\alpha = .26$$

This means that the nominal student will be incorrectly classified in about one case out of every four on the average, given the particular sampling plan chosen in this example.

Evaluating the merits of this particular sampling plan leads to the conclusion that the economy afforded by having only three items is offset to some extent by the fact that, of the pupils who actually meet design standards, over 25% are likely to be improperly classified as nonmasters. The important question this raises is the relationship between test length and efficiency. For example, by adding one or two more questions to the test, by how much could the efficiency be improved? To find the answer to this question, we proceed to a simpler and more general formulation of the probability space than the probability measure than that used in the above example.

3.42 Scoring Formulas and Data Reduction

The fact that certain test outcomes have the same probability as seen from Table 3-4, suggests that it is possible to reduce the data in some way without losing any useful information. A transformation which effects this reduction is the application of a scoring formula (p. 80). If we let $u_g = 0,1$ denote the binary response to an item and

$$\underset{\sim}{v}_a = (u_{1a}, u_{2a}, \ldots, u_{n_k a})$$

be pupil $\underline{a}$'s item response vector to $\underline{n}_k$ items, then the scoring formula can be written as a simple function of $\underset{\sim}{v}_a$:

$$x_a = x\,(\underset{\sim}{v}_a) = \sum_{g=1}^{n_k} u_{ga} \qquad (3\text{-}2)$$

The possible test outcomes expressed in terms of observed scores, x_a, for the previous 3-item example are shown in the bottom row of Table 3-4. The fact that no information is lost in this data reduction procedure, other than the order in which right and wrong responses occur, is illustrative of what is meant by calling x_a a sufficient statistic. The reason why no information is lost can be found first in the assumption of local independence and secondly in the random selection procedure. The order of the responses is not important because it is assumed that every response is independent of every other response and every item is instructionally as important as every other item. Thus a reordering of a given set of items would not be expected to change the probability of a correct response to an item and therefore it would not change the probability of a given test score, x_a, occurring.

The use of a scoring formula to achieve data reduction results in both syntactic and semantic changes in the formulation of the probability space (Ω,P). At the semantic level, the points, or elementary outcomes, in Ω which were item-response patterns in the case of the quantal response model become test-scores in the case of the item-sampling model. In the former case, Ω consists of 2^n distinct points for an n-item test. In the latter, this number is reduced to $(n + 1)$ points, the number of possible test scores.

The new probability measure defined on this set of $(n + 1)$ points represents the corresponding syntactic change in the model and is derived from the item-sampling model as follows. From Table 3-4 we see that any given score, x_a, occurs only if one of the mutually exclusive response patterns given in the column above the score-value is observed for pupil a. Therefore the probability that a score, $X = x_a$, will be observed is computed by using the "sum" rule. Employing the item-sampling model rather than the quantal response model, these probabilities can be expressed in the form, $\mathrm{Prob}(X(\underset{\sim}{V}) = x(\underset{\sim}{v}) \mid \zeta)$, for the 3-item illustrative test as follows:

$$\mathrm{Prob}\ (X = 3 \mid \zeta) = \zeta^3\ ;$$
$$\mathrm{Prob}\ (X = 2 \mid \zeta) = 3\zeta^2 \cdot (1 - \zeta)\ ;$$
$$\mathrm{Prob}\ (X = 1 \mid \zeta) = 3\ \zeta \cdot (1 - \zeta)^2\ ;$$
$$\mathrm{Prob}\ (X = 0 \mid \zeta) = (1 - \zeta)^3\ ;$$

The method of generalizing this probability measure for an arbitrary n-item CRT is evident by noting that the coefficients on each line are computed by counting the number of combinations in which x successes can chose from among n possible outcomes. The familiar computing formula for this is the binomial coefficient:

$$(3\text{-}4) \qquad \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Thus, in general, we have for an n-item CRT that

$$\mathrm{Prob}\ (X(V) = x(v) \mid \zeta) = f(x_a)$$

where

$$(3\text{-}5) \qquad f(x_a) = \binom{n}{x_a} \zeta_a^{x_a} (1 - \zeta_a)^{n-x_a}$$

and

$$x_a \in \Omega = \{0,1,2,\ldots,n_k\} .$$

3.43 Lattice Diagrams, CRT Performance Paths, and Criterion Boundaries.

In order to investigate the relationship between efficiency, test length, and criterion value for a given sampling plan using the item-sampling model, it is convenient to visualize the sample space, Ω, in terms of a "lattice diagram" such as that shown in Figure 3, Appendix G.

A pupil taking a test can be regarded as starting at point (0,0) on the lattice diagram and stepping 1 unit to the right for each correct response and 1 unit upward for each wrong response. In this way he completes a random walk which terminates at one of the lattice-points on the line,

(3-6) $x + w = n$

where w is the number of errors for pupil a on the test.

The lattice points on the line (3-6) may be considered as images of the points in Ω. This particular view is especially helpful in deriving the operating characteristic of single sampling and curtailed sampling plans.

Figure 4, Appendix G, shows a more complete lattice diagram for the case n=20 on which have been drawn an illustrative random walk together with two criterion-lines. We shall call the horizontal line, $w_1 = c$, the error criterion. This means, for pupil a, if $n-x_a < c$, that he has met the test criterion. If, on the other hand, $n-x_a \geq c$, then the pupil has failed to meet the criterion. A vertical "success" criterion line is shown at the point $x_c = n-c+1$. The performance path for pupils who meet criterion will intersect this line.

Using the lattice diagram with criterion boundaries drawn on it, one can see that the image points on the line (3-6) which represent an "accepted lot" lie beneath the level of the error criterion line, $w_1 = c$, while those points which signify a "rejected lot" lie to the left of the vertical (success) criterion line. The random walk, which we shall call the pupil's performance path, suggests a number of ideas with regard to acceptance sampling which will be needed shortly.

So far we have referred to test efficiency mainly in terms of Type I errors. This kind of error is indicated when the performance path for a "master" pupil touches the line $w_1 = c$. The probability of this event occurring is equal to the chance that the number of errors the pupil makes is in the range $c \leq w \leq n$. The maximum probability of such a classification error occurring is suffered by the nominal student, i.e. one whose proficiency $\zeta = \zeta_1$. Therefore, using (3-5) and the sum rule, we have

$$\alpha = \sum_{w=c}^{n} \binom{n}{w} \zeta_1^{n-w} \zeta_1^{\prime w} : \tag{3-7}$$

The opposite kind of error can also occur, i.e. the path for a pupil whose proficiency is actually below the nominal limit could cross the line $x_c = n - c + 1$. The cause of errors of either kind is, as noted earlier, due to random environmental and personal factors that we call "noise." However, the consideration of errors of Type II raises a problem of defining a "nominal nonmaster," a matter which will be deferred to the next section.

More importantly, we can project the possibility of shortening the test by observing the random walk on the lattice-diagram. As soon as the path touches either criterion boundary, the decision to be made in regard to the test outcome is completely determined. Therefore, if it would be possible to detect this event economically, one could reduce the cost and time needed to reach prescriptive classification decisions on the basis of CRT data. This possibility will be explored more fully in the sections on the CSSP and the SPRT.

3.5 Operating Characteristic of a Single Sampling Plan (SSP).

A test which runs to completion, in the sense that the terminal point of the imagined random walk lies on the line (3-6) is called a single sampling plan of a size n or "SSP." An SSP is *defined* when one selects the value of test length, n, and the (error) criterion value, c.

A sampling plan of size n and criterion c has a certain predictable operating characteristic for sorting pupils into mastery and nonmastery groups. This operating characteristic, or OC, as it is called in the sampling inspection literature (SRG, 1945; Wald, 1947: Wetherill, 1966) is simply the probability that the student's score will meet criterion. From (3-5) and the addition rule of probabilities for mutually exclusive events, the OC for a SSP is

$$\text{(3-8)} \quad S = \text{Prob}\ (w < c\ ) = \sum_{w=0}^{c-1} \binom{n}{w} \zeta^{n-w} (1-\zeta)^{w}.$$

That is, the probability of a successful criterion performance, S, equals the sum of the probabilities of that fewer than c errors will occur in n item-response trials, given a proficiency ζ .

This probability of success, S, is effectively illustrated if we plot S as a function of the independent variable, ζ. In so doing, we obtain a curve that looks remarkably like a classical item-characteristic curve. Figures 5 to 7, Appendix G, show the OC's computed for SSP's of $n_k = 5$, 20, and 25 respectively with error criterion values in the range $c = 1, 2, \ldots, (0.40\ n_k)$. The curves shown in Figures 5, 6 and 7 are called the operating characteristic curves, or OC-curves, for a Single Sampling Plan.

3.51 Ideal Sampling Plan Characteristics.

The operating characteristic of a sampling plan offers a method of specifying acceptance requirements which conform as closely as desired to the general type of design requirements described in Table 3-1. It should be noted that the validity of the scheme to be suggested depends on the extent to which the derivation of the OC is valid and that this in turn hinges upon the degree to which the criterion-theory of Chapter Two models the real world. Given that the model is adequate in its first-order representation, the elements of an instructional quality control management decision system are as follows.

Suppose first that we have a design requirement for instructional process k which specifies that a learner should possess a minimal proficiency $\zeta < 1.00$ with respect to SCO_k. The specification of the corresponding acceptance requirement that determines whether, in fact, this design requirement is met requires first that the "product" the learner submits for test be sampled in some fashion and secondly that a decision rule be given by which one decides that the sample either does or does not meet design requirements.

The single sampling plan (SSP) requires only that a random sample of size n be drawn from the SCO. We shall see how to specify n shortly. An ideal decision rule would have the following characteristics. For any pupil having proficiency $\zeta \geq \zeta_c$, it should be decided that his learning product is up to design requirements; for any pupil having proficiency $\zeta < \zeta_c$, it should ideally be decided that his learning product does not meet design specifications.

The OC for such an idealized acceptance requirement would be a step function such as that shown in Figure 8, Appendix G. Note that the abscissa for this and subsequent OC-curves is scaled in terms of error rate, $\zeta' = 1 - \zeta$, rather than proportion correct, ζ, for the reasons indicated earlier.

However, the acceptance requirements for the plan idealized in Figure 8 cannot actually be written since the decision rule assumes no error of measurement. Since, as we have shown earlier, error always occurs in any measure of a continuous quantity, which proficiency is assumed to be, there is no possible method of measuring ζ in such a way as to always be able perfectly to separate proficiences at and above a given point from those below that point.

As a step toward more realistic acceptance requirements, it seems plausible to separate the alternative regions of decision by a range of proficiency which represents a region of indifference. For example the following decision rules might be established.

(1) If $\zeta' \leq \zeta'_1$, accept H_0.

(2) If $\zeta' \geq \zeta'_2$, accept H_1.

The range $0 \leq \zeta' \leq \zeta'_1$, is the critical region for accepting H_0. The range $\zeta'_2 \leq \zeta' \leq 1.00$ is the critical region for accepting H_1. The limit, ζ'_1, is the maximum error rate that limits the region of proficiency definitely considered to meet design requirements.

The step which introduces a region of indifference requires that a second proficiency criterion be specified in the design requirements in addition to ζ'_1. Call this second proficiency ζ'_2 and define it to be the least error rate definitely considered to fail to meet design requirements. The corresponding proficiency, ζ_2, is what we shall call the "nominal nonmaster" proficiency. By imposing the condition

$$\zeta'_2 > \zeta'_1$$

we have a nonempty region $\Delta \zeta = \zeta'_2 - \zeta'_1$ for which we are indifferent whether to decide in favor of H_0 or H_1.

The OC-curve for this specification of the acceptance requirement shown in Figure 9, Appendix G. Again the problem of infinite sampling is encountered because the error of measurement is assumed to be zero at each of the two criterion points.

3.52 Admissable Sampling Plan Characteristics

The crucial step to take to reduce n to a reasonable value for classroom testing lies in permitting the possibility of erroneous decisions to arise from the acceptance procedure. The decision strategy, to be practical, must permit the possibility of deciding incorrectly that design requirements have not been met by a pupil's learning product when in fact it meets specifications and, conversely, that the product is accepted when indeed it fails to meet design specification. The four possibilities of two kinds of possible correct decisions and two kinds of possible incorrect decisions are shown in Figure 10, Appendix G. As is customary in hypothesis testing, we take α to denote the maximum relative frequencies at which errors of the first kind may occur and β to denote the maximum tolerable limit for errors of the second kind. Then $(1 - \alpha)$ denotes the level of confidence for a given decision shceme and $(1 - \beta)$ the power.

With these concepts, it is possible to specify practical acceptance requirements for the learning product of a given instructional process. Analytically, one derives β in the same way that the (3-7) was derived.

The maximum probability for an error of type II occurs for nominal nonmasters, i.e., when $\zeta = \zeta_2$ and such a pupil makes fewer than c errors:

$$\text{(3-9)} \qquad \beta = \sum_{w=0}^{c-1} \binom{n}{w} \zeta_2^{n-w} \zeta_2'^{w}$$

The correct values of n and c which specify acceptance requirements giving levels of protection α and β against erroneous decisions are obtained by solving equations (3-7) and (3-9) simultaneously for n and c in terms of ζ'_1, ζ'_2, α, and β. Since this is not a straightforward computational problem, what often can be done is to reverse the procedure by selecting n and c so that approximate values of α and β obtain at proficiency levels ζ_1 and ζ_2.

3.53 Empirical Considerations for Selecting An SSP.

An empirical procedure to follow, for example, might utilize concepts similar to the industrial notions of average quality limit (AQL) and lot tolerance percent defective (LTPD). The AQL corresponds to the specification of ζ'_1 and the LTPD to the specification of ζ'_2. Customary values of α and β associated with the AQL are 5% and, with the LTPd, 10%. Thus, for example, Figure 5 shows that a test of 5 items with error criterion $c = 1$ will erroneously reject about 5% of those learning products submitted for test by pupils with about 99% proficiency (AQL corresponds to $\zeta'_1 = .01$) and, conversely, erroneously accept about 10% of the samples tested for pupils with 62% proficiency (LTPD corresponds to $\zeta'_2 = .38$). The region of indifference in this example is the proficiencies in the range $.01 < \zeta' < .38$.

Similarly in the case where the acceptance requirements specify $n = 5$ and $c = 2$, the 5% AQL corresponds to an error rate of about

$\zeta'_1 = 8\%$ while the 10% LTPD corresponds to an error rate of about $\zeta'_2 =$ 55%. The region of indifference in this case is $.08 < \zeta' < .55$.

Inspection of curves in Figures 5, 6 and 7 indicate the effects of changing *n* and *c* in the acceptance requirements. An increase in *n* reduces the size of the region of indifference. The OC reflects this by showing a steeper decline for sampling plans with larger n. Increasing the value of the error criterion *c* for fixed *n* is equivalent to increasing the both values of the error criteria ζ'_1 and ζ'_2 which limit the critical regions of decision.

Table 3-5 shows representative values for various values of *n* and *c* together with a corresponding set of values for α, β, ζ', and ζ'_2. There are many ways of considering the selection of *n* and *c* depending on how one imagines the design requirements to be specified. In any event, it is clear that the not unusual practice of setting fixed percentage criteria for tests of variable length really represents quite different acceptance requirements when specified in terms of α, β, ζ'_1, and ζ'_2. For example, suppose one chose the fixed percentage criterion to be 80% for tests of any length whatever (e.g. Coulson, *et al* 1968). For the sake of comparison, consider two "equivalent" 80 per cent criterion plans denoted by ($n = 5$, $c = 2$) and ($n = 20$, $c = 5$). The average quality limit is slightly lower in the $n = 20$ case while the LTPD is considerably higher. If the objective of such acceptance specifications is to correctly classify all those pupils having 80% proficiencies or better, it is evident by reading the graphs

TABLE 3-5

Comparison Data on Acceptance Requirements

Test length, A =	5		20								25	
Error crit. c =	1	2	1	2	3	4	5	6	7	8	1	10
(ζ = 5%) ζ'_1 =	.01	.08	.005	.04	.06	.08	0.11	0.15	0.12	0.22	.001	.25
(β = 10%) ζ'_2 =	.38	.55	.13	.18	.24	.31	.37	.42	.48	.53	.09	.51
$\Delta\zeta = \zeta'_2 - \zeta'$	.37	.47	.13	.19	.18	.23	.26	.27	.30	.31	.09	.26
$\zeta' \ni \partial s/\partial\zeta'$ is max.	1.00	0.70	.98	.94	.90	.86	.82	.78	.73	.69	1.00	.65

in Figures 5 and 6 that the 20 item test misclassifies only about 3% (α) of such pupils while the 5 item test gives much poorer protection with $\alpha = 28\%$, approximately.

If one wishes to discriminate between levels of achievement represented by smaller regions of indifference, the longer the test must be made. The selection of the error criterion, for fixed *n*, can be estimated as the value of ζ' for which $\partial s/\partial\zeta'$ is an absolute maximum, about the midpoint of the region of indifference (see Table 3-5). Finally, for specified levels of protection against erroneous classification at given points ζ'_1 and ζ'_2 on the proficiency continuum, it is incorrect to set a fixed percentage criterion c as a function of test length, *n*. The efficiency of the test varies considerably with n, becoming very inefficient for small *n*. Protection can be approximately maintained at a fixed level at one criterion point only for variable *n* and *c*.

3.54 Procedures for Selecting an SSP.

The selection of *n* and *c* for a CRT can be regarded as being analogous or functionally equivalent to the selection of item-β and X50 for items on an NRT. Therefore one should choose the minimum value for *n* that provides adequate discrimination together with a value for *c* that centers the region of discrimination between the nominal proficiency limits given in the design specifications. Such a procedure is very similar to the technique, familiar in industrial quality control sampling, of specifying an average quality limit for the product as well as a lot tolerance percent defective.

The selection of n and c has been found to be most practically effected, in the pilot experiments run to date, by inspection of OC curves such as those shown in Figures 5, 6, and 7. OC curves have been tabulated and are readily available for a wide variety of sampling plans (e.g. SRG, 1945). In addition, Appendix D includes the listing of two computer programs which will generate the OC for any of the types of sampling plans described in this paper.

3.6 Curtailing the Single Sampling Plan

Various cost considerations indicated earlier suggest a need to reduce the number of items in a CRT to the minimum needed for adequate decision-making. If we refer to the lattice diagram of Figure 4, it is evident that once the performance path crosses either the failure boundary (w=c) or the success boundary (x=n-c+1), the outcome of the test is determined. The principal problem lies in the cost of setting up a sampling system which detects the signal that the performance path has met a criterion boundary.

The most promising technology for curtailing tests appears to be in the use of teletype terminals under computer control. Objective costs of such a system involve a fixed overhead for leasing the terminals plus a variable cost which is proportional to the number of items required for decision-making. The essential question is whether the reduction in number of items needed through the use of curtailed tests will save enough time and money to offset the cost of terminal service. The following is an analysis of curtailed single sampling plans which provides a method for determining the answer to such a question.

Two matters of interest in regard to curtailed sampling plans are (1) the operating characteristic, as in the case of the single sampling plan, and (2) an additional concept, the average test length of the curtailed plan. The operating characteristic must, of course, be the same for the CSSP as that of the single sampling plan inasmuch as the final decisions whether to accept H_0 or H_1 are identical in either case. However, the test length, which is fixed at some value $n = N$ for the SSP, is a random variable in the case of the CSSP.

Although the test length is unpredictable in the case of curtailed tests, it is of interest to calculate the average test length or ASN (for Average Sample Number). Expressions for the operating characteristic and for the average test length are derived from the assumptions of the item-sampling model for criterion-referenced tests, as follows.

Consider the pupil's test performance path on the lattice diagram of Figure 4. The test continues as long as this path does not touch the criterion boundaries. If the performance path touches the error criterion boundary, the test terminates and we know that the pupil has made exactly c errors. However, the number of correct responses may lie anywhere in the range from 0 to $n-c$, where n = the maximum test length. Therefore, the test length, in the case where the test terminates in rejection of H_0, is a random variable lying between the limits c and n.

Similarly, if the test terminates by having the pupil's performance path cross the success boundary, $x_c = n-c+1$, then we know that $(n - c+1)$ is the exact number of correct responses made whereas the number of errors may have ranged from 0 to no more than $(c-1)$, by definition of error criterion. By noting that the average test length is mathematically defined as the expected value of the random variable n, we may calculate the ASN for a curtailed test. This is done by first computing the probability that the test will terminate after n questions have been asked, then multiplying each possible value of n by its probability of occurrence, and finally summing all these products to get the expected value or average.

The sample space for curtailed tests can be visualized on the lattice diagram as the set of lattice points on the two criterion lines. To compute the probability of a particular outcome when H_0 is rejected, we first note that the final response must be the error which produces the last vertical step in the performance path which brings it up to meet the criterion line. The total number of questions asked may be as few as c, if all errors are made at the outset, or as many as n, the maximum possible limit. Let m be an index that denotes the number of possible correct responses $(m = 0, 1, 2, \ldots, n-c)$. Then the probability of a test of length n occurring when H_0 is rejected is the product of (1) the number of ways to distribute the $(c-1)$ errors remaining among the $(c+m-1)$ possible opportunities for error by (2) the probabilities of exactly $\underline{c}$ errors and $\underline{m}$ correct respones being observed:

$$(3\text{-}10) \qquad \Pr(\underline{n} = c+m) = \begin{pmatrix} c+m-1 \\ c-1 \end{pmatrix} \zeta^{m} \zeta'^{c}$$

A similar derivation follows for the case where the test ends in the acceptance of H_0. In this event, the final response must be correct in order that the performance path terminate with a horizontal step into the success criterion line. The test record thus consists of exactly n-c+1 correct responses and a variable number of errors ranging between 0 and a maximum of (c-1). Let the index m denote the number of possible errors, in this case (m = 0, 1,...,c-1). Then the probability that a test of length $\underline{n}$ is observed when H_0 is accepted is the product of (1) the number of ways in which the $\underline{m}$ errors are distributed among the (n-c+m) possible opportunities for error by (2) the probabilities of $\underline{m}$ errors and (n-c+1) correct responses occuring:

$$(3\text{-}11) \qquad \Pr(\underline{n} = n-c+1+m) = \begin{pmatrix} n-c+m \\ m \end{pmatrix} . \zeta^{n-c+1} . \zeta'^{m}$$

Equation (3-11) is the OC for the curtailed sampling plan, i.e. the probability of product acceptance given the quality, ζ. Given (3-10) and (3-11), we are now able to calculate the expected test length of Average Sample Number (ASN) according to the definition

$$(3\text{-}12) \qquad E(n|\zeta) = \sum_{m=0}^{n-c} (c+m) \begin{pmatrix} c+m-1 \\ c-1 \end{pmatrix} \zeta^{m} \zeta'^{c} +$$

$$\sum_{m=0}^{c-1} (n-c+1+m) \begin{pmatrix} n-c+m \\ m \end{pmatrix} \zeta^{n-c+1} . \zeta'^{m}$$

Figures 11, 12, and 13, Appendix G, show ASN's for curtailed plans which have operating characteristics identical to those shown in Figures 5, 6, and 7, Appendix G.

Inspection of these curves shows that significant reduction in test length is possible over the SSP with no loss of efficiency. However, the SSP requires no special means of test administration other than conventional paper and pencil materials whereas the CSSP requires that one be able to make a decision following each response. Such interactive testing seems most effectively done through the use of remote terminals such as the teletype, thus adding a cost to the CSSP which the SSP does not incur.

3.7 Decision-Theoretic Considerations in Sampling Plan Selection

From the consideration of the single sample and the curtailed sampling plans, it is evident that acceptance requirements can be written in a variety of ways, with each associated sampling plan having its own advantages and disadvantages. This raises interesting questions concerning the selection of an optimum plan from the set of available plans. As indicated earlier, utility theory is designed to solve this kind of problem. We sketch here an overview of the decision-theoretic selection of sampling plans before going on to discuss a curtailed sampling plan that has certain optimum properties, Wald's SPRT.

First, we will need to refer to certain definitions, commonly used in decision theory (these have been adapted with suitable modification from Wetherill, 1966):

Loss refers to costs averaged over the set of outcomes of any sampling plan for a given value of the error rate parameter, ζ':

Risk refers to costs averaged over a prior distribution of ζ'.

The cost of testing is proportional to the number of items in the test. We shall take as a unit the cost of testing one item.

Two cost functions need to be know in addition to the cost associatd with the ASN of a selected sampling plan. These are

$W_0(\zeta')$ = cost of accepting H_0 when a pupil's error rate is ζ'.

$W_1(\zeta')$ = cost of rejecting H_0 when the error rate is ζ'.

Let

S denote a particular sampling plan.

and let

$P_A(\zeta'|S)$ = probability of accepting H_0 under S when the error rate is ζ'.

$P_R(\zeta'|S)$ = probability of rejecting H_0 under S when the error rate is ζ'.

Thus $P_R = 1 - P_A$. Now let

$E(n|\zeta',S)$ = ASN, the expected sample size, given ζ' and S.

Then the loss function can be defined as

$$R(\zeta'|S) = E(n|\zeta',S) + P_A(\zeta'|S)W_0(\zeta\ ) + P_R(\zeta'|S)W_1(\zeta').$$

From an aconomic standpoint, the "best" sampling plan for an SCO is the plan S' such that

(3-13) $R(\zeta'|S') < R(\zeta'|S)$ for all available plans, S.

If such sampling plan S' exists such that (3-13) holds for all values of ζ', then it is called an optimum plan.

Two principles for selecting a sampling plan when no optimum plan can be found are the following. In general, the risk $R(\zeta'|S)$ varies both with ζ' and S. For any given sampling plan, there will be some value or set of values of ζ' for which $R(\zeta'|S)$ is a maximum. However, this maximum cost will probably be different for different plans, S. The minimax principle is used to select the plan S'' which minimizes this maximum loss:

Minimax Principle: Choose

$$S'' = \underset{S}{\text{Min}} \left(\underset{\zeta'}{\text{Max}}\, R(\zeta'|S)\right)$$

Such a plan S'' is called an admissable plan. It can be seen that the use of the minimax principle is essentially the most pessimistic basis for selection of a plan since it assumes the worst possible value of ζ' will occur.

A modification yields a selection principle that is less pessimistic, called the principle of minimax regret. Assume as before that the selection of a plan S''' involves a cost for each value of ζ', $R(\zeta'|S''')$. There may be other plans which involve less loss for this particular value of ζ'. The difference between the cost $R(\zeta'|S''')$ and the least cost possible under any other plan S is called the regret. That is,

$$U(S) ::= \text{Regret} = R(\zeta'|S) - \min_{S} R(\zeta'|S)$$

The amount of "regret" will in general vary with ζ'. The minimax-regret principle bases the selection of a sampling plan on the following rule which minimizes the maximum regret:

Minimax Regret Principle: Select S''' such that

$$U(S''') = \max_{\zeta'} (R(\zeta'|S''') - \min_{S} R(\zeta'|S))$$

is a minimum.

If a process curve or prior distribution is known concerning the frequency with which CRT measured proficiencies, ζ, occur, then the expected loss

$$E(R(\zeta'|S))$$

can formally be minimized with respect to S.

3.8 The Sequential Ratio Probability Test (SPRT).

In general, analytic techniques for selecting sampling plans are cumbersome and not particularly effective. However, there is one well-known optimum sampling plan that may have considerable potential as computer technology is made available to the classroom. The plan in question is based on sequential analysis as developed by Wald (1947) and the Statistics Research Group (1946). Technologically, the requirements for using the SPRT in educational testing appear to be similar to those needed for administering a CSSP. Basically, one must have the capacity to administer tests via an interactive terminal.

In brief, the SPRT decides after each response (1) whether the pupil's proficiency, ζ, is at or above the nominal mastery level, ζ_1, or (2) whether his proficiency is at or below the nominal nonmastery level, ζ_2, or (3) whether no decision can be made, in which case another item is presented for inspection. In order to apply the test, one must know in advance the distribution function for the random variable, ζ, under consideration. By operating with the assumptions of the CRT model, the appropriate distribution to use is given by (3-5), the binomial.

The SPRT operates in much the same manner as the CSSP but with the following unique decision rules. Following each trial question, the "likelihood ratio" is computed:

$$(3\text{-}14) \qquad L = \frac{\zeta_2^{x}(1-\zeta_2)^{t-x}}{\zeta_1^{x}(1-\zeta_1)^{t-x}}$$

where t = the number of trials and x = the number of correct responses observed in the _t_ trials.

The following rules are then applied after each item response:

a. If $L \leq \frac{\beta}{1-\alpha}$, accept H_0.

(3-15) b. If $L \geq \frac{1-\beta}{\alpha}$, accept H_1.

c. If $\frac{\beta}{1-\alpha} < L < \frac{1-\beta}{\alpha}$, continue testing.

From (3-14) and (3-15), it is clear that the acceptance requirements for the SPRT are determined, not by specifying _n_ and _c_, as in the case of the SSP and CSSP, but by specifying the quadruple $(\zeta_1, \zeta_2, \alpha, \beta)$.

The OC and ASN for the SPRT are computed from relatively complicated formulas compared to those for the CSSP. Pilot experiments indicate that the additional cost, as measured in computer time, is essentially the same for both plans when the OC's are constrained to be nearly identical. This implies simply that the additional calculation required by (3-14) and (3-15) is negligible. Formulas used to construct the computer program listed in Appendix D are briefly summarized in Appendix E.

In general, it is possible to specify SPRT acceptance requirements so that the OC is nearly identical to that of a given CSSP. The essential question that remains is, given that the two plans are equally efficient, which involves least cost? Figure 14, Appendix G, compares the ASN for a CSSP defined by the pair (n=5,c=2) with both the fixed test length for the corresponding SSP and the ASN for a SPRT, all plans having nearly the same OC. Inspection of the graph reveals that the SPRT effects a savings for all values of ζ when compared with the other plans, except $\zeta' = 0$, in this particular case of interest.

One of the practical problems in implementing the SPRT in a school testing situation has been uncovered in pilot experiments. The SPRT decision process is theoretically infinite although the probability of termination increases rapidly with $\underline{t}$. A number of procedures exist for terminating the sampling inspection by some maximum value of $\underline{t}$ (e.g. Amster,but Wald and Wolfowitz (1948) have shown that the SPRT is an optimum plan only when the quality of the product is equal to either the nominal limit for mastery and for nonmastery. This implies that one must be careful in stating the acceptance requirements for the SPRT so that not many pupils tested are likely to have proficiencies which lie in the region of indifference. If such an adverse situation were met, the cost of administering the SPRT can exceed that of a comparable CSSP by an intolerable amount.

3.9 Summary Comparison of Curtailed and Single Sampling.

Curtailed plans compare disadvantageously with the SSP in the following important ways:

1. The CSSP requires a decision to be made after the administration of each item.
2. Since test lengths vary, there may be some administrative difficulties due to some children finishing before others if a group administration is employed.
3. Curtailed samples provide a poorer estimate of ζ' for the child's behavior on an SCO.

However, there are the following advantages:

1. Since ASN $\leq$ N, where ASN is the average sample number for a CSSP and N is the sample size of a SSP, there can be appreciable savings in test time, thus enabling one to "manage" more SCO's, as well as save in testing costs.
2. Machine decisions are not publicly made, thus the embarrassment or sense of failure that often occurs in conventional hand operated group tests may be alleviated.

Curtailed sampling plans are not feasible unless one has computer-assistance in the form of interactive terminals which can present the items and accept the response. This poses some advantages as well as disadvantages beyond those mentioned above when the CSSP is compared with single sampling plan.

The simplest and perhaps most economical system would require little computational capacity and thus would enable one to employ the least expensive of today's generation of time-sharing computers. The principle computer requirement would be effective input/output (I/O) data handling capabilities. Tests could be stored in memory administered, curtailed, and interpreted on the basis of the decision algorithms discussed previously. Records of student performance would be kept by the machine; recommendations to the pupil for further study could be obtained by a computer talbe-lookup routine which would find descriptions of available resources or prescriptions associated with failure to achieve a specified criterion; and elementary data reduction could be performed on accumulated records to enable the machine to generate summary reports to teachers or pupils as needed. The details of this reporting aspect have been explored and demonstrated in several educational settings and need not be expanded in detail here (e.g. Suppes _et al._, 1968; Coulson _et al._, 1968).

Basically the same procedures can, of course, be used in connection with single sampling plans. However, in the SSP case, no important use is made of the terminal's interactive capacity or of the machine's capacity for applying decision rules after each response. One would probably do better to have tests printed and kept on file rather than to administer the tests via teletype. Following an administration of any test, data transmission to a computer would likely be more economical by utilizing machine readable answer sheets. An optical scanner could then be used to convert the responses into computer usable form and, if desired, transmitted in a batch to the computer effecting some savings in line time compared to TTY terminal operation. In this case, computer outputs could be returned on a line printer or, if small amounts of data were involved, less expensive hard copy devices might be employed.

Both the CSSP and SSP can be used in group testing situations by applying the decision rules for stratified proficiency groups described earlier. Actually one is treating proficiency groups as though each group were a single individual. However, the CRT permits one to evaluate, diagnose, and prescribe on an individual basis if that is desired. This would be the case, for example, in a continuous progress environment. In such a situation, it is the CRT's property of providing absolute rather than group-mean referenced measures that makes individual decision-making possible in a relatively simple sort of way.

The essential requirement for managing individual programs in a continuous progress learning environment would be to know proficiency at some initial point in time (e.g. from a pretest score) and, from knowledge of the individual's learning characteristics, to have some approximate estimate of his proficiency following his learning experience. This data is needed to select an admissable sampling plan to measure his proficiency and to output information, following the posttest, concerning his absolute level of attainment. Diagnostic information and suggested prescriptions might be based on an analysis of items missed or on which the latency of response was large. Table lookup procedures could also be employed, of course. In this regard, further study is needed to identify learner characteristics which are predictive of achievement or proficiency gains associated with given instructional treatments.

The preparation of fixed tests which are stored and recalled from computer memory is the usual way the assessment problem has been handled in the past, usually in connection with an SSP on which inadequate consideration was given to the choice of $\underline{N}$ and $\underline{c}$. One could improve the performance of such management systems as have been proposed by utilizing the procedures discussed herein to create item pools and administer samples from these pools according to an admissible sampling plan. But there is another possibility that warrants further study, namely the machine generation of random samples of items from a given pool using the rules which define the SCO.

Machine generation of items suggests that each pupil may receive a different sample than that given to any of his peers. The machine uses the generation rules to compute the desired response, outputs the stimulus or question, measures the latency of response, records the response input by the pupil, evaluates it by formula, phonetic, or character-by-character comparison rules, decides whether to curtail or not, and upon termination of the test outputs appropriate instructional management data. Appendix B contains samples of tests generated in such manner by the computer program listed in Appendix F.

Such a procedure would effectively use the strengths of the computer. It seems likely, for instance, that item generation and possibly even prescription generation would have more practical potential than existing lookup schemes for extending the number of terminals and the number of SCO's that an instructional management system could handle. For this reason, it is urgent that further research be undertaken to develop practical methods of computer generation of items as well as computer generation of prescriptions.

Obviously, chronic problems of test security would be significantly eased by such an approach. Possibly important savings in testing cost may be effected also. The matter needs to be studied carefully, comparing the costs and relative advantages of management systems operating in SSP batch mode and those utilizing interactive terminals and CSSP procedures.

CHAPTER IV

IMPLICATIONS FOR CONTINUED PROGRAMMATIC RESEARCH

4.0 Overview.

Implications of this study for the development of better educational practice can conveniently be treated in the following four areas. First, the Criterion-Referenced Test theory and the associated management system are, in some respects, radically different from certain traditional approaches to educational measurement and management. Therefore, the theory needs to be tested and validated in several ways. Some implications regarding questions that need to be answered and methods for seeking these answers are discussed in Section 4.1.

Secondly, the new methods of instructional management which have been described have many implications for improving curriculum and instruction, possibly at all levels, but especially at the K-12 school levels. Validation of hypothesized curriculum hierarchies and the measurement of the effectiveness of competitive instructional methods are examples of two problems given particular attention in Section 4.2.

Thirdly, there has been discussion of a systems approach to education in recent years which, for lack of a clear theoretical basis, has not always been substantive. The theory and methods of management described in this dissertation have implications for putting the systems approach to education on a sound theoretical and technological basis.

Implications for a possible educational systems discipline are treated in Section 4.3.

Left to the last, but prossibly most important, are the implications for teacher education. Topics dealt with the Section 4.4 relative to this area include (1) applying techniques for instructional management to the management of teacher education, (2) preparing teachers to assume differentiated roles in inquiry-orienteated schools, and (3) developing modern professional competencies in educational workers.

Educational research has been criticized over the years on several accounts, many of which stem from its apparent irrelevance to the practical needs of the classroom. The means provided by the management and measurement techniques described in this paper have implications particularly on this point of applicable research. Some relevant questions and implications for research in education are discussed in connection with each of the above-mentioned areas of concern.

4.1 Implications for Further Development of Criterion-Referenced Test Theory.

Francis Bacon (e.g. 1960 was perhaps the first to recognize that the validation of a scientific theory must proceed in what is essentially a negative fashion. There is no way to prove, in a positive or deductive sense, that a theory is true. The scientific method involves, rather, a process of seeking and testing hypotheses which if rejected would cause the theory to fall. The demonstration of positive results does not, on the other hand, establish a theory any more than exhibiting positive instances can establish the truth of a mathematical theorem.

Thus, it has been observed that a theory which cannot be rejected is not a theory at all since it is, under such circumstances, capable of embracing contradictory conclusions. In order to validate a theory or model, therefore, it is essential to identify the means of answering two basic questions: (1) which hypotheses can one test to reject the theory and (2) what hypotheses does the theory in question reject (Platt, 1963)?

The answer to such questions results in a form or level of validation that Bruner (1965) and others have called formal validation. At another level, it is possible to lend credence to a model by showing that it, in fact, does work in the real world. Such a method, therefore, might be described as functional validation. Basically, functional validation consists of determining the real world limitations and situations under which the assumptions of the model hold.

A theory can be supported also at yet another level which might be called the affective level. Here the essential ingredient is a judgmental or intuitive feeling that the theory or model is "right" or sensible. One looks for the degree to which implications of the theory conform with reasonable expectations. Although validation at the affective level is likely to be more variable and therefore less reliable than functional or formal validation, intuitive judgment is nevertheless an important adjunct to consider together with the other means of testing a model. In the following remarks, then, implications for the validation of the test and management models proposed in this paper are considered at the affective, functional, and formal level. It should be noted here that the significance of validation efforts may lie not so much in whether a particular model stands or falls but in the new insights and the progress toward better rechniques that is made possible by the various validation processes.

4.11 Affective Validation.

This CRT Test and Management Theory implicitly takes issue with the prevalent feeling that scientific or systematic methods can come about in the classroom only through the use of some pyschological learning theory. Behavioral learning theory is the current popular focus, of course, but the issue is not with any one particular learning theory. The truth is that we know little about how one learns which can be put to reliable use by the classroom teacher. It is in this hard fact that one reason can be found for the failure of much educational research to change events in the classroom.

The CRT theory described in this paper makes no assumption of a learning-theoretic nature with the exception perhaps of the implications of local independence. A consequence of this point must be, therefore, that systematic objectives and methods of instruction can be drawn up independently of learning theory. Antithetically, however, it cannot be concluded that learning theory itself is either devalued or invalidated; only that it is at best a sufficient but not a necessary concomitant of a functioning instructional management system.

If this is so, then efforts to cast all instructional activities into a behavioral mold may be ill-advised. A better way may lie in the following direction. Even a cursory inspection of the way in which teachers and text-writers behave, in the absence of viable knowledge of how one learns, reveals that in subject matter areas like mathematics and reading, where an underlying discipline exists, problems or stimuli or exercises are grouped into classes in which the similarity of the items is quite obvious. In other words, the notion which has been formalized in this paper as a specified content objective is manifest in the intuitive instructional behavior of teachers and textbook authors. Thus the CRT model seems right or promising to the extent that it implies an approach to instruction that is compatible with observed practical behavior on the part of experienced educational workers. One of CRT-theory's principal contributions lies simply in the fact that it provides the teacher or author a method of determining efficient sample sizes on an analytical rather than intuitive basis. This should help correct the present tendency to run toward one extreme or the other in the number of items employed, both in providing practice and in testing for achievement.

While these observations yield a degree of affective validation to CRT theory itself, they more importantly imply that continued study should be given to fine optimum ways to organize SCO's and to find out, for example, how many SCO's need to be identified for successful instructional management. These are in a sense the standard questions of curriculum scope and sequence recast in terms of the systems approach developed in this paper.

It is not likely that present methods will reveal a mathematically determinable optimum number of SCO's. Yet one can estimate that the number of manageable objectives for one subject area in any given semester will not exceed 100 by even one order of magnitude. This implies that the degree of specificity with which one identifies SCO's need not be so refined as to lead to thousands of objectives such as was done in the days of Social Utility Theory and is yet being done in some behaviorally-oriented curriculum design work today. Rather, CRT theory implies the need to get a feel, through the process of systematic trial and revision, for a practical way of organizing instruction about objectives that can be pre- and post-tested in the available time and at reasonable cost. Convergent testing techniques seem to be called for and the CRT model seems well-suited to guide such investigation. In particular, the minimum-number-of-questions property of CRT sampling plans together with the use of item-generation techniques provide promising tools for investigation of curriculum structures especially in individualized settings.

For example, Figure 15 (Appendix G) shows a 20-item exercise generated by a computer. The generation rules for this SCO define the "homogeneous" population of all 2x2-digit multiplication problems. Nevertheless, inspection of the sample reveals considerable item diversity. For example, some items have the smaller operand at the top while others have it placed in the second position, a matter only of chance occurrence. Some items (e.g. #2 and #8) require "basic" or tens multiplication; others do not. Some require much regrouping or "carrying" (e.g. #18) while others such as #2 require very little.

A traditional view would judge the items to be of greatly varying difficulty, which they no doubt are, and therefore not homogeneous items. However, homogeneity can be specified with respect to many kinds of parameters of which item p-value is only one. In the case of individualized testing, an item that is difficult for one student of interest may not be difficult for another. In other words, the concept of item difficulty lacks universality and applicability in the essential situations with which the instructional manager is confronted.

Teachers instinctively recognize this and instead check samples of a given class of problems, not with special concern for statistical properties of the items, but rather with attention to the patterns of error that appear. Where such patterns are observed, diagnosis and prescriptive treatment can then be made.

In general, the classes of items found in test and practice samples are not p-homogeneous. Nor would that be an ideal situation, since patterns reflecting the degree of concept and skill attainment would probably be less easy to identify. Rather, the items are what might be called zeta-homogeneous or homogeneous with respect to the proficiency which is built into the design specifications associated with the instructional package. Thus they are homogeneous also with respect to the instructional objectives from which item-generation rules are drawn. While many different learning concepts and behaviors may be implicit in the requirements of the various items in such a pool, nevertheless all the requirements are related to the SCO, such as the 2x2 digit problems of the above example, and therefore, possess a kind of unity that is useful to the instructional manager.

If a student exhibits a given level of proficiency on such samples of items, it is a necessary consequence that he will be at least as proficient on all component skills and concepts. Thus one CRT may measure lower bounds of proficiency on a set of several subskills. Some care must be taken to include enough subskills on one SCO to make it possible to include all the desired manageable areas into the 100 or so SCOs which define instructional goals for a semester. But if too many are grouped together, other problems can arise, such as increased difficulty in diagnosis. An illustration of the problem that can arise is provided by those conventional tests of mathematics achievement which divide items into very coarse classifications, such as arithmetic items, geometry items, spatial perception items, etc. The patterns of error become too complex to identify readily and prescriptive quality degrades with the degree of diagnostic resolution. Such instruments are, in general, ineffective for instructional management purposes.

Thus, it seems that an important implication stemming from affective or intuitive considerations is that organization of the curriculum might fruitfully be approached in a non-behavioral, yet systematic fashion. The emphasis would be on getting a manageable number of objectives sufficiently well-defined to permit quick and accurate diagnosis rather than on generating pools of questions having certain kinds of statistical item-properties relating to person-populations, such as p-value, item-beta and X50. All this implies a new direction for research, not so much toward the optimally sequenced curriculum structures of CAI, but rather toward the design of better strand/unit block structures which have been found to be effective in inquiry-oriented systems of instruction emphasizing self-selection and self-pacing principles.

CRT theory can provide an effective basis for organizing this kind of research since it brings the researcher rather quickly to consider two practical ramifications of his speculations: (1) it must be possible to design item-generation rules or item populations that are instructionally homogeneous with respect to hypothesized content objectives and (2) it must be possible to coalasce highly specific categories of objectives into larger, more manageable specified categories. The latter suggests a mechanism for organizing objectives into strands and units according to their logical relationship rather than on the basis of a strictly hierarchical relationship. Ideally the strands, then, would be mutually exclusive and exhaustive of the subject-matter universe as would the units also be with respect to the strands treated as the universe.

4.12 Functional Validation.

Appendix F contains copies of the computer program, record files, and sample outputs for a prototypic model of an instructional management system which deals with units from several strands in the typical arithemtic curriculum. Choice of the particular segment for experimental development was based largely on the immediate needs of the cooperating school. It represents no particular set of constraints or limitations in itself on the applicability of CRT theory. Yet it does show in miniature the basic features of a functioning management system as envisioned here.

The program operates in five possible modes described as follows. Mode 0 is a non-interactive mode for producing master copies of tests and practice items, together with a key; items are randomly generated from a STANDARD population. This population is parametrically specified by the records entered into the file called STAN/CMS or, alternatively, when Test #0 is requested, the program permits the teacher or teacher-aide to enter a subset of the control variables to produce a customized set of items. Records in STAN/CMS are 80-column card images. Each record is defined by its sequential position in the file and accessed by entering the test number corresponding to this position upon request of the program. Parameter values are preset in the file by using a file-handling systems program called EDIT. Parameters being used in the present form of the program are defined in the following format:

OPCDUPL1-LOL1-UPL2-LOL2-UPL3----LOL3----CBONIHVABRMDMSCRZ1Z2ALBTBA

where

OPCD::= a 4-digit field for the specification of an operation; (1=addition; 2=subtraction; 3=multiplication; 4=division); a single value in the first position followed by zeroes produces items all involving the same operation; if more than one non-zero value is entered, in any order, the program randomly selects one of the legal operations specified before generating each item. Test #20 (Appendix F) is an example of this capability where all four operations are sampled.

UPL1-::= a 5-digit field specifying the upper limit of the first operand.

UPL2- and LOL2-::= as above, but applying to the second operand.

UPL3- and LOL3-::= as above, but applying to the result.

CB::= a 2-digit field controlling the regrouping requirements of each item; (-1=control off; 00=no regrouping in any item; 01=regrouping required in every item);

O::= a 1-digit field specifying the number of operands in a column addition problem (2 to 9).

NI::= number of items for SSP, 2-digit field (01 to 99).

HV::= problem format control, 2-digit field; (-1=random selection of horizontal or vertical format; 00=always horizontal format; 01=always vertical format).

A,B::= two 1-digit fields for specifying the place value of the quotient at which zeroes are to appear, counting left from the unit's place.

RM::= 2-digit field which determines if division problems have remainders or not (-1=remainders randomly occur; 00=remainders never occur; 01=remainders always occur).

D,M::= two 1-digit fields for specifying the number of digits in the multiplicand and multiplier when carry/borrow option is invoked.

S::= sampling plan selector (0=SSP; 2=SPRT).

CR::= fixed error criterion for CSSP operations.

Z1,Z2,AL,BT::= Four 2 digit fields which define a SPRT sampling plan; (ZL=AQL; Z2=LTPD; AL=level of significance; BT= probability of a Type II error).

BA::= numeration system selector (base two to ten currently operable).

Characteristics of other program modes are as follows:

MODE 1:= Flexible Interactive Mode; options include, among others, the choice of items randomly selected from a STANDARD pool or from a custom-specified pool; choice of either vertical or horizontal format or random selection of either format; correct-answer echo control (OFF for testing, ON for practice); SSP, CSSP, or SPRT acceptance sampling decision-rules may be applied as desired; pupil records sorted into files named GIESEM/CMS (for pupils exhibiting mastery performance levels) and GIESEN/CMS (for those failing to meet acceptance requirements). GIESE/CMS is a file containing records of all interactive events.

MODE 2:= Production Interactive Testing Mode. Item-response vectors recorded in file ITEMS/CMS for automatic sorting into homogeneous proficiency groups. Test items are randomly sampled for the first test but remain the same for all examinees in a given group. Mode 2 is useful for analyzing pre-test data particularly. Item response vectors can also be entered by paper-tape techniques from hand-scored tests produced in Mode 0 for rapid instructional-grouping service based on pre-test data. Mode 2 is always used as a test and never a practice mode.

MODE 3:= Dummy Mode. Permits entry of a previously recorded starting value for the random number generator in order to secure a copy of an earlier test or practice set. Starting values are recorded in the various output files with suffix "CMS," mentioned above. This mode essentially enables the compact memorization of all stimuli generated any given sequence by recording one 8-digit number.

MODE 4:= Like Mode 2, a production interactive mode. Overhead time is reduced to entering only pupil names after the initial selection of test and other option settings. Mode 4 differs from Mode 2 in that all pupils in the group receive different random samples drawn from a SCO. Also the SSP, CSSP, and SPRT acceptance sampling techniques apply in Mode 4 whereas Mode 2 employs an experimental technique still under development. Mode 4 can be used for either test (answer echo OFF) or practice (echo ON) purposes.

Experimentation with this measurement and management system to date provide support for the CRT model at the functional level. Exemplary data, in the sense of being typical rather than carefully selected, is shown in the summary forms of Appendix A. It is interesting to note the relatively high coefficient-α (reliability) values for short (5-item) CRT's.

This initial work has pointed to certain problems that will require further attention in the near future. At a mechanical level, there will be the need to determine protocols for using the computer service in the most effective way. Mode 0 provides the most economical service level; Mode 1 is most expensive; Modes 2 and 4 lie somewhere in between.

Using the available Burroughs-5500 information processing system, conservative cost estimates indicate that, for approximately $5 per school day per 100 pupils, it will be possible to maintain proficiency profiles where the instructional organization includes a combination of teacher-directed and self-directed learning modes. Traditional classwork and individually guided inquiry have been mixed together in pilot tryouts of the System, the ratio of time spent by a given pupil in each mode depending on the pupil's developing capacity for self-directed learning. At an increased cost amounting to approximately $10/school day/100 pupils, it appears that such additional services as selective remedial drill and practice, pre-test analysis for instructional grouping, diagnostic test analysis, and individual pupil prescription service can also be provided.

Thus initial cost estimates indicate reasonable and effective service can be obtained for approximately the cost of a teacher's salary for the academic year. However, there are many variables associated with each mode which will be useful to explore. For example, it has been found feasible to make a teletype (TTY) copy of a test or drill sheet and use heat-transfer duplicating techniques to produce multiple copies for pupils. Experimental system operation has utilized a teacher aide to operate the TTY remote computer terminal to secure desired test or practice forms, and to score and return the results to the computer files by hand. The use of paper tape to economize on the telephone line time should be a high priority item for consideration in this phase of the operation. Also in this connection, a half-duplex mode has been used to date, which increases the risk of undetected errors in data communication due to ambient electrical and acoustical noise levels. The economics and technical feasibility of full duplex operation should be explored to eliminate this problem.

Another step to be taken in developing the system would be in the direction of multi-school management of intermediate mathematics programs. Extensions of the computer program, MATH/CMS (Appendix F), to include the universes of real and rational numbers, the symbolism of generalized open sentences and additional numeration systems, inequality relations, and certain structural properties of familiar number systems are readily foreseen. By developing existing and projected services in the form of multi-user programs, one would have an important vehicle for investigating the economies of management system dissemination across schools and grade-levels.

More importantly, the roles of teacher and aide require attention. There appears to be a prestige factor associated with operation of the TTY which enhances the aide's status and may detract from the teacher's status both in the eyes of the pupils and among the staff. This arises now largely because of special knowledge the aide has gained concerning computer operation while the teachers have not had time nor a functional need to learn these things. The effects are therefore largely psychological and therefore of considerable importance. Possible solutions include teacher-inservice to provide a comfortable operating knowledge of the computer system. At the root of the matter, however, there is a need to provide teachers with a perception of the potential that computer-management has so that they can contribute to the development of the System. Only with an adequate grasp of the potentials and limitations of computer-assistance will it be possible for teachers to display the leadership needed to develop individual learning modes and to fill the role of knowledgeable decision-makers that their professional status demands in such situations. It seems appropriate, therefore, to consider implications of this need in the area of teacher education; particularly that more attention be given to inquiry methods as a part of the prospective teacher's own education and to developing in the teacher a growing familiarity with modern educational technology.

The following provides an illustration of the kind of judgment required in using the CMS system. It has been found that the testing features of the interactive modes have been useful in measure response latencies. This information appears to be especially valuable at the basic skills level. However, various uses for the interactive modes remain to be explored. Of particular interest is the selection of practical SPRT sampling plans. Initial plans selected for test have been borrowed largely from industrial protocols which typically set $\alpha = .05$, $\beta = .10$. It quickly becomes apparent that this selection favors the "producer's" interest at the expense of the "consumer."

That is, it seems preferable to the producer to run higher risks of accepting bad lots (since this cost passes on to the consumer) while minimizing the risk of rejecting good lots (which raises the cost to the producer). From an instructional point of view, it may be more desirable to reverse these values since "rejection" implies that help will be provided the pupil to bring certain concepts that he goes on to subsequent learning activities. Acceptable balance points which require reasonable sample sizes remain to be determined through systematic trial and revision. Decision-theoretic tools may also be applicable to this problem.

The CRT-management system was originally conceived to assist in managing self-paced, self-selective, individual-inquiry systems of instruction in school mathematics. The prototypic model exhibited in Appendix F marks a point of progress at which concern can now turn from the most fundamental matters of system organization to the next cycle of development. Using data handling capabilities of the existing system, the following areas should be explored.

Advisory Prescriptions. Previous experience with IMCP has yielded some information how children successfully select subsequent strands and units for study. However, one obstacle has been that self-selection processes have often been pre-empted by teachers who have difficulty in separating their instructional practice from traditional directive techniques. This behavior in turn seems to be prompted by a fear that effective control will be lost over the learning process. However, with the assistance of computer-management this fear should be eased and serious study of the self-selection process may begin. An empirical Bayesian approach seems one of several likely methods of approaching this kind of study. Recently developed techniques for analyzing nominal scale data (e.g. Press & Roger, 1967) also may be useful.

Strand-Unit Reorganization. Existing strands and units were designed prior to the conception of the computer-management system. Recent work on developing a prototypic system for one complete strand has indicated some ways of improving the strand/unit organization. As indicated in the previous section, this matter can be handled more efficiently than before by using CRT-theory rather than learning theory as a practical organizing focus.

Item-Generation. Item-generation techniques are essential to consider in any system requiring frequent testing. Parallel tests appear to be most economically generated using techniques such as those demonstrated by the program in Appendix F. Currently restricted to the domain of whole numbers, it would be a straightforward task to extend the available universes to include integers, decimal fractions, common fractions, mixed numbers, real numbers, and complex numbers as needed. The existing system treats numerical symbolism in bases two to ten. More elaborate programs for bases larger than ten or for ancient systems of numeration could be devised.

The program should also be extended to include inverse operations by providing for spaceholder positioning in any of the three possible locations, as desired, in horizontal format. Relations, other than equality, are natural extensions to consider. It may be desirable to extend the number of operations beyond the basic four, but this is not urgent. Finally, thought should be given to find ways to implement the structural properties of the various number systems in order to get at SCO's strictly at the concept level. Also, it is not too soon to consider extending these techniques to the reading area, especially at the primary level.

Hardware. Full duplex operation and paper-tape capabilities have been mentioned earlier. Due to the noise of the TTY and the rapid development of terminal technology, some investigation of soft-copy devices and optical scanning systems seems warranted as a next step. Costs are likely to be relatively high but experience so far has indicated several methods of economizing. In particular, manufacturers of such terminals should be encouraged to market less, rather than more sophisticated hardware for particular application to instructional management, where the needs are different from mass processing of NRT test data. Steady, day-by-day CRT-management requires more modest hardware requirements than does high-surge twice-a-year or four-times-a-year norm-testing. Finally some improvement in the accuracy of latency measurements is needed. A hardware interrogation function to determine the moment when the TTY becomes write-ready is available on the B-5500 System for this purpose.

4.13 Formal Validation.

CRT instruments, like any test, must detect true variance in proficiency among pupils when variance exists. Reliability is a measure of this test property. It should be relatively easy to test the hypothesis that CRT's can have high reliability. If this hypothesis fails, CRT theory as proposed here, clearly fails. The tests in Appendix A give positive evidence in support of the high-reliability hypothesis. Further exploration of the CRT's statistical properties should be conducted, however.

The concept of proficiency introduced in Chapter II implies another hypothesis that is subject to direct test. If two or more proficiencies compose some global proficiency (e.g. basic multiplication and column addition proficiency are needed for 2x2 digit multiplication), then the CRT model implies that the global proficiency should be the product of the component proficiencies. This hypothesis should be tested in many settings. If support is found for the hypothesis in some cases, but not others, there may be reason to rethink whether or not the hypothesized component skills are in fact components of a given global skill. In other words, once the principle in question appears to have adequate support to accept its validity, it may be turned around, so to speak, to assist in task-analysis research. Rejection of the hypothesis can occur not only due to the fact that it may be false but, if it is in fact true, due to the failure to identify all significant subtasks or the inclusion of irrelevant subtasks in the hypothesized composition of the global task.

Finally, there is one significant circumstance under which an otherwise adequate CRT's reliability coefficient may become very small or even negative. This occurs when the group of examinees is ζ-homogeneous. If true ζ-variance is negligible, the item-intercorrelations, KR-20, and external-criterion correlation drop to mean-sero values. This expected property has been detected in work to date, as shown in Appendix A. Should it fail to occur, some fault with the theory would be implied.

Here again, there is a possibility of turning traditional statistical applications around for instructional management purposes. Empirical sorting of pupils into nearly homogeneous proficiency groups can be effected by utilizing a Mode 2 type of procedure. If a selected sub-group has little true variance in proficiency, then KR-20 calculated for the sub-group alone should be near zero. Hi-values of KR-20 would suggest resorting to achieve less score variance. Eventually, one would expect to find ζ-homogeneous groups using the test reliability statistic as a sorting index. This technique should be given further study. Research completed to date indicates that 5-item tests may be used to identify up to 3 homogeneous sub-groups. If further study bears this out as a general property of short CRT's, one could further test the assumptions of the first-order model by using, for example, Kolmogorov-Smirnov to test the expected binomial for goodness-of-fit. Bimodal distributions could appear in some instances, indicating the presence of second-order processes within an SCO. If such evidence is formed, there would be reason to extend the model to account for second-order phenomena. Conversely, one could split the SCO in some way to achieve greater homogeneity in the items in order to conform to the first-order model.

4.2 Implications for Curriculum and Instruction.

4.21 Identifying Curriculum Hierarchies.

As noted earlier, MacDonald (1965) has identified the rationalist "myth" as one guide to curriculum construction which is currently centered in the mainstream of educational thought. The CRT-system developed here certainly is consistent with the rationalist myth, as contrasted with developmental, aesthetic, or moralistic myths. Thus, CRT-theory is in the tradition begun by Tyler and Herrick. However, there is major point of division on the matter of behaviorism. The rational approach to curriculum construction is marked by its insistence on setting rational objectives, then designing instructional segments to achieve the objectives, and finally evaluating the learning product to see if the objectives were obtained. However, in recent years, this approach has come to be based almost exclusively on the behaviorist approach identified with such as Skinner, Gagné, Glaser, Silbermann, Ammons, Lumsdaine, Walbesser, etc. CRT-theory is properly viewed as a viable alternative to this approach but still within the same philosophical framework. SCO's and design requirements take the place of behavioral objectives; CRT's and absolute acceptance requirements take the place of evaluation by norm-referenced relative ranking. And while behaviorism attempts to apply corrective and sequencing mechanisms by means of reinforcement and extinction techniques applied on system command, CRT-management is designed to suggest modifications in the instructional groupings and the self-selection sequence on demand as pupil or teacher may request prescriptive and diagnostic assistance.

Given this background, the concept of a curriculum hierarchy is not a necessary precondition to a functioning CRT-managed curriculum as it is in most CAI systems. Nevertheless, observation of natural self-selection processes in an inquiry-oriented demand-system may yield information useful to behavioral command systems of curriculum design. Hierarchical relationships can be sought in the context of many possible parameters since the ordering must be based on some observable dimension, e.g. difficulty. The free inquiry mode characteristic of IMCP may reveal that hierarchical structures can be associated with particular parameters which define populations of students. This is a phase of research which the CRT-management system makes increasingly attractive.

4.22 Predictive Learning Theory.

While many studies have been conducted on the nature of the learning curve, they have not dealt directly with the development of a school child's academic proficiency, in the sense used here. If we assume that there is some sort of relationship between instructional effectiveness and proficiency development, we might proceed--along the lines of Carroll's learning model (1963)-to hypothesize certain properties concerning the relationship which suggests the form for the learning curve.

If we think of a graph of proficiency against time, then ideally the proficiency curve should sweep upward, rapidly at first, finally curling over to gradually approach some limit. In other words, proficiency should be a bounded, increasing function of time. In the simplest case, ignoring what we might call the "rust" problem, we could assume a monotonically increasing relationship between proficiency and time. That is, we ignore the fact that proficiencies decay or "get rusty" over the time through disuse.

In order to determine the position and shape of an ideal proficiency curve, we note that at absolute time, t = 0, academic proficiency can be approximated at $\zeta_{ak} = 0$. Thus the curve should pass through the origin, (0,0). Now let t_0 stand for the starting time of some instructional segment and let ζ_0 indicate the proficiency that the student has already developed before the instructional treatment begins.

Suppose there exists some mediating factor called instructional effectiveness, denoted by ϕ, operating in a learning situation. As a measure of instructional effectiveness, ϕ would have the following properties. "Better"instruction would be characterized by a ϕ-value higher than that associated with "worse" instruction and ϕ would affect the slope of the learning curve, particularly. That is, whatever function specifies the growth of proficiency as a function of time should also show that growth is a function of instruction.

One way the form of such a relation might be induced is as follows. To begin with, it does not seem reasonable to assume human beings to be capable, in general, of displaying 100 per cent proficiency. Very likely the maximum potential proficiency individuals might be capable of is a random variable with an approximately normal distribution, a characteristic of "ability" distribution. Assume, in any case, that some non-perfect upper limit to proficiency exists for each pupil.

Let C_{ak} represent the upper limit to the possible proficiency development for individual #a on SCO #k. Since C_{ak} is a proficiency, it follows that the range of values for C_{ak} must be $0 \leq C_{ak} < 1$.

It seems reasonable to hypothesize that a proficiency growth curve would initially increase rapidly with time since the rudimentary proficiency levels ususlly seem easier to develop. As time goes on, the expected gain for a given period of time would likely decrease as ζ asymptotically approaches the limit C_{ak}.

A function that generally fits all these requirements is given by equation 4-1, the exponential function:

$$(4\text{-}1) \quad \zeta_{ak}(t) = = C_{ak}(1-e^{-\phi_{ak}\,(t+t_0)})$$

Equation (4-1) is plotted for arbitrary values of 0 and t in Figure 16 (Appendix G). It may be noted that the subscripts a and k are needed to account for the individual effects of instruction.

Certain interesting properties of (4-1) can be drived by considering some sample values. We not first that when $t = {}^{-}t_0$ (which marks the beginning point of proficiency development), we have:

$$\begin{aligned} \zeta({}^{-}t_0) &= C_{ak}\ (1-e^{-\phi(t_0+t_0)}) \\ &= C_{ak}\ (1-e^{0}) \\ &= C_{ak}\ (1-1) \\ &= 0 \end{aligned}$$

Thus the curve goes through the origin when the time axis is scaled in absolute terms.

Figure 16 shows the time scale shifted by the constant t_0 so that time is measured from a relative zero, the time at which an instructional treatment begins. Thus when t= 0 (relative) we have:

$$\zeta_0 = \zeta(0) = C_{ak}\ (1-e^{-\phi(0+t_0)})$$
$$= C_{ak}\ (1-e^{-\phi t_0}),$$

a constant that is characteristic of the individual pupil. ζ_0 can be interpreted as the value of proficiency which we attempt to estimate on a pretest. It is the proficiency the child already has developed before instruction begins.

Then as $t \longrightarrow \infty$, the quantity $e^{-\phi(t + t_0)}$ gradually approaches zero. Since this is an amount subtracted from 1, ζ (t) gradually approches $C_{ak} \cdot (1 - 0)$, or simply C_{ak}.

Given these properties, we next solve 4-1 for instructional effectiveness, ϕ, to study its developing relationship with the independent variables. First, dividing (4-1) through by C_{ak}, we get

(4-2) $$\frac{\zeta(t)}{C_{ik}} \approx \frac{\Delta\zeta}{C_{ik}} = 1-e^{-\phi(t+t_0)}, \text{ (assuming } \zeta_0 \text{ is small).}$$

The ratio on the left side of the equation (4-2) compares the gain in proficiency since the beginning of instruction, $\Delta\zeta$, with C_{ak}. In effect, the absolute proficiency gain, $\Delta\zeta$, is converted in this step to a relative measure based on the maximum gain possible for a given individual (which would be $\Delta\zeta = C_{ak} - 0 = C_{ak}$). The relation (4-2) can be interpreted as "evening-out" comparisons between students experiencing a common instructional treatment. For example, a student characterized by $C_{ak} = 1$ would have to make a gain of 0.75 in order to have the same relative measure of proficiency gain as one whose $C_{ak} = 0.4$ and who showed a proficiency gain of 0.3. Thus the step (4-2) can be interpreted as an adjustment for the individual's "achievement" ($\Delta\zeta$) due to instruction relative to "ability" (C_{ak}) before arriving at an evaluation of the effectiveness of instruction.

The next step in solving 4-1 for ϕ involves transposition of two terms to yield:

$$(4\text{-}3) \qquad e^{-\phi(t+t_0)} = 1 - \frac{\zeta(t)}{C_{ak}} = \frac{C_{ak} - \zeta(t)}{C_{ak}}$$

The quantity on the right side of (4-3) has the appearance of a probability complement. Thus the quantity on the left might be interpreted as a measure of the relative amount of undeveloped proficiency remaining at the time t following the start of instruction.

The next steps in solving for ϕ require first taking the reciprocal of each member and then the natural logarithm, thus:

$$(4\text{-}4) \qquad \phi \cdot (t+t_0) = \ln\left(\frac{C_{ak}}{C_{ak} - \zeta(t)}\right)$$

Now let t_f denote the duration of instruction (Carroll's "available time") and $\Delta\zeta = \zeta\,(t_f)$ be the proficiency gain attributable to the instructional treatment. Then

$$(4\text{-}5) \qquad \phi = \left(\frac{1}{t_f + t_0}\right)\ln\left(\frac{C_{ak}}{C_{ak} - \Delta\zeta}\right)$$

$$(4\text{-}6) \qquad \phi = \ln\left(\frac{C_{ak}}{C_{ak} - \Delta\zeta}\right) \frac{1}{t_f + t_0}$$

From (4-6) we see that $\phi = 0$ when

$$(t_f + t_0)\sqrt{\frac{C_{ak}}{C_{ak} - \Delta\zeta}} = 1$$

or when

$$\frac{C_{ak}}{C_{ak} - \Delta\zeta} = 1 \Rightarrow C_{ak} = C_{ak} - \Delta\zeta \Rightarrow \Delta\zeta = 0$$

Thus we have $\phi = 0$ when $\Delta\zeta = 0$.

This implies that if instruction results in no change in proficiency, then the measure of instructional effectiveness is zero, as desired. At the other extreme, if proficiency developed as a consequence of instruction is the maximum possible, i.e. $\Delta\zeta = C_{ak}$, then ϕ is infinite.

Thus we find that the range of the index ϕ, is $0 \leq \phi \leq \infty$. No natural unit appears in the derivation for since instructional effectiveness is measured as a ratio of like quantities. Hence it is a "pure" number.

It is interesting to note that the parameter, ϕ , not only indicates an objective measure of instructional effectiveness but the relation (4-5) bears a striking resemblance to one of the forms of the so-called Weber-Fechner Law, in particular to Fechner's Massformel. The reader will recall the Weber-Fechner law provided the first real, though limited, opportunity to measure unobservable levels of human sensation by relating it to a measure of the observable stimulus in units called the "just noticeable difference" or jnd. $\text{Eq}^{\underline{n}}$ (4-6) suggests that it may be possible, in a similar fashion to obtain approximate measures of instructional (or prescriptive) effectiveness in units of just noticeable differences in observable proficiency. It, therefore, implies some unique possibilities for applying the CRT theory of this paper to measure the relationship between learning and teaching.

One related implication for further study is the use of the computer to build instructional simulation models to study empirically the distribution of students being taught according to some specified instructional plan (e.g. the 3-aptitude group plan which is characteristic of Los Angeles schools). Valuable information concerning system stability, reliability, and efficiency might be generated by such a study. As a cost estimating device, such simulations might be valuable to run before an actual large-scale IMS were put into the schools on a trial run. Somewhat similar kinds of simulation studies have been reported (Cogswell, 1964) but none take an instructional effectiveness parameter into account explicitly.

4.23 Implications for Assessing Instructional Effectiveness

The parameter ϕ may be useful in the analysis of instructional situations of more modest proportions. This has always been a very sensitive area to research because of the natrually deep ego-involvement of teachers with this kind of study. Since the equation (4-6) helps separate important aspects of instructional effectiveness from possibly confounding effects, such as time allotted to instruction and initial levels of proficiency, it may be possible to design more clear-cut and less emotion-packed studies through the use of CRT theory and the model suggested by (4-6) then has previously been the case. Hopefully (4-6) will stimulate the composition of competing models and alternative hypotheses. These may, in turn, lead to a fuller understanding of measureable instructional components.

If (4-6) or some other proficiency curve can be found to describe the growth of proficiency and hence learning, it may be possible to relate instruction more effectively to the characteristics of the learner. For example, if instructional packages of known ϕ were available, relative to a specified target population, one could use a relation such as (4-6) to predict how long it would take an individual to achieve a given level of proficiency. Alternatively, (4-6) could be used to predict the level of proficiency that would be expected as the result of a fixes period of instruction.

Among other things (4-6) implies that Carroll's (1963) concept of aptitude as the period of time required to achieve mastery should be considered a function of ϕ and not solely some natural or nurtural characteristic of the learner.

That CRT results are indicants of instructional effectiveness can be seen in the summary shown in Appendix A. The transition matrices indicate the quantity of success, measured in numbers of pupils who make the nonmaster—> master transition during instruction. The sub-group proficiency levels measure the quality of learning in terms of the absolute level of proficiency attained. Therefore, if one had several competing instructional methods, their relative effectiveness might be estimated by means of such data. This was the kind of significant result Hammock (1960) had in mind for urging that further work be undertaken to develop CRT theory. The results in Appendix A tend to support his argument for its effectiveness in this regard.

4.3 Implications for the Systems Approach to Education.

The systems approach to education, or at least the label, is less than a decade old. A survey of headings in the Education Index shows that the term evolved in connection with the impact of computer technology in programmed instruction. In the late Fifties, colleges were being urged to include courses in systems or "operations" analysis. Computers were still classified as calculating machines, and automation was considered by some as a means to alleviate the teacher shortage. Programmed Instruction and CAI seem to have reached their zenith in the early Sixties after which the idea of automated teaching seemed to generate a counter reaction. The systems approach, with all its promise for putting economy and order into the relatively chaotic and uneconomically-oriented educational enterprise, became increasingly a catch-phrase. The literature today shows the systems approach to be little more than empty exercises in flow-charting, a sort of 2-dimensional means of replacing the conventional method of outlining programs in linear form.

The CRT-theory developed in this study was in no small way motivated by a desire to search for substance in systems technology that could be applied to educational problems. Some success can be claimed on this count and further research can be recommended. Primarily, research is needed to identify admissable sampling plans for gathering decision/diagnostic/prescriptive data. This study has produced evidence to indicate that rarely will more than 25 or fewer than 5 items from an SCO be needed to make good instructional decisions. But this still leaves a virtual infinity of ways to specify sampling plans. Multiple sampling has not been touched in this paper. Nor has the application of these same techniques to process control problems been more than mentioned. Each of these appear now to be potentially promising areas for developing a substantive educational systems theory and technology. The area is large enough, risky enough, and important enough to dedicate energies on the order of an entire Research and Development Laboratory to it. The work described in this dissertation is but a small step in this direction.

4.4 Implications for Teacher Education.

At no level is an individualized, inquiry-oriented approach to education more desirable than at the college and university level. In particular, if teachers are to learn how to teach or manage classes in an inquiry mode, they certainly should experience the basic techniques applied to their own education. It therefore seems appropriate that attention should be given to the use of computer-assisted management techniques in connection with teacher education.

It is commonly believed that beginning teachers teach much as they have been taught. If this is true, it is essential that the freedom to explore and inquire without experienceing ensuing chaos, which computer management makes possible, be incorporated as a part of a teacher's experience at the earliest, most formative stages.

In addition, a new era requires that teachers be aware of the possible role of computers as well as its limitations. Treated as an important tool to free the teacher to be more effective in small group situations, the computer deserves to be a part of the teacher-preparation curriculum.

Successful development of computer-managed systems of instruction has many implications for all involved in the educational enterprise. Successful small group instruction techniques would have a revolutionary impact on teacher education. The principle of differentiated elementary school staffs would gain viability as would the need for many new classifications of educational technologists.

Implied is the need for course work in computer technology, computer-assisted instruction, and even in computer programming for teachers, aides, curriculum analysts, instructional designers, and so on. Importantly, viable alternatives to the presently narrow focus on behaviorism may provide the basis for teaching teachers how to organize instruction in the context of design and acceptance specifications, sampling techniques, and systematic decision-making. Grading practices, long condemned for their harmful judgmental effects, could be replaced by regular reports that would resemble itemized accounts of SCO's attempted and proficiency levels attained. In other words, it may be possible to approach more closely the long-sought ideal of teachers, pupils, and resources interacting fully and inquisitively in a search for knowledge unhampered by the anachronistic and repressive administrative constraints that still operate in too many of today's schools.